

Universal and Composite Hypothesis Testing via Mismatched Divergence

Jayakrishnan Unnikrishnan, Dayu Huang,
Sean Meyn, Amit Surana and Venugopal Veeravalli

Abstract

For the *universal* hypothesis testing problem, where the goal is to decide between the known null hypothesis distribution and some other unknown distribution, Hoeffding proposed a universal test in the nineteen sixties. Hoeffding's universal test statistic can be written in terms of Kullback-Leibler (K-L) divergence between the empirical distribution of the observations and the null hypothesis distribution. In this paper a modification of Hoeffding's test is considered based on a relaxation of the K-L divergence, referred to as the mismatched divergence. The resulting mismatched test is shown to be a generalized likelihood-ratio test (GLRT) for the case where the alternate distribution lies in a parametric family of distributions characterized by a finite dimensional parameter, i.e., it is a solution to the corresponding *composite* hypothesis testing problem. For certain choices of the alternate distribution, it is shown that both the Hoeffding test and the mismatched test have the same asymptotic performance in terms of error exponents. A consequence of this result is that the GLRT is optimal in differentiating a particular distribution from others in an exponential family. It is also shown that the mismatched test has a significant advantage over the Hoeffding test in terms of finite sample size performance for applications involving large alphabet distributions. This advantage is due to the difference in the asymptotic variances of the two test statistics under the null hypothesis.

Amit Surana is with United Technologies Research Center, 411 Silver Lane, E. Hartford, CT. Email: SuranaA@utrc.utc.com. The remaining authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL. Email: {junnikr2, dhuang8, meyn, vvv}@illinois.edu.

This research was partially supported by NSF under grant CCF 07-29031 and by UTRC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or UTRC.

Portions of the results presented here were published in abridged form in [1].

Keywords: Generalized Likelihood-Ratio Test, Hypothesis testing, Kullback–Leibler information, Online detection

I. INTRODUCTION AND BACKGROUND

This paper is concerned with the following hypothesis testing problem: Suppose that the observations $\mathbf{Z} = \{Z_t : t = 1, \dots\}$ form an i.i.d. sequence evolving on a set of cardinality N , denoted by $Z = \{z_1, z_2, \dots, z_N\}$. Based on observations of this sequence we wish to decide if the marginal distribution of the observations is a given distribution π^0 , or some other distribution π^1 that is either unknown or known only to belong to a certain class of distributions. When the observations have distribution π^0 we say that the *null hypothesis* is true, and when the observations have some other distribution π^1 we say that the *alternate hypothesis* is true.

A decision rule is characterized by a *sequence* of tests $\phi := \{\phi_n : n \geq 1\}$, where $\phi_n : Z^n \mapsto \{0, 1\}$ with Z^n representing the n -th order Cartesian-product of Z . The decision based on the first n elements of the observation sequence is given by $\phi_n(Z_1, Z_2, \dots, Z_n)$, where $\phi_n = 0$ represents a decision in favor of accepting π^0 as the true marginal distribution.

The set of probability measures on Z is denoted $\mathcal{P}(Z)$. The relative entropy (or Kullback-Leibler divergence) between two distributions $\nu^1, \nu^2 \in \mathcal{P}(Z)$ is denoted $D(\nu^1 \parallel \nu^2)$, and for a given $\mu \in \mathcal{P}(Z)$ and $\eta > 0$ the *divergence ball* of *radius* η around μ is defined as,

$$\mathcal{Q}_\eta(\mu) := \{\nu \in \mathcal{P}(Z) : D(\nu \parallel \mu) < \eta\}. \quad (1)$$

The empirical distribution or *type* of the finite set of observations (Z_1, Z_2, \dots, Z_n) is a random variable Γ^n taking values in $\mathcal{P}(Z)$:

$$\Gamma^n(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{Z_i = z\}, \quad z \in Z \quad (2)$$

where \mathbb{I} denotes the indicator function.

In the general universal hypothesis testing problem, the null distribution π^0 is known exactly, but no prior information is available regarding the alternate distribution π^1 . Hoeffding proposed in [2] a generalized likelihood-ratio test (GLRT) for the universal hypothesis testing problem, in which the alternate distribution π^1 is unrestricted — it is an arbitrary distribution in $\mathcal{P}(Z)$, the set of probability distributions on Z . Hoeffding's test sequence is given by,

$$\phi_n^H = \mathbb{I}\left\{ \sup_{\pi^1 \in \mathcal{P}(Z)} \frac{1}{n} \sum_{i=1}^n \log \frac{\pi^1(Z_i)}{\pi^0(Z_i)} \geq \eta \right\} \quad (3)$$

It is easy to see that the *Hoeffding test* (3) can be rewritten as follows:

$$\begin{aligned}
\phi_n^H &= \mathbb{I}\left\{\frac{1}{n} \sum_{i=1}^n \log \frac{\Gamma^n(Z_i)}{\pi^0(Z_i)} \geq \eta\right\} \\
&= \mathbb{I}\left\{\sum_{z \in \mathcal{Z}} \Gamma^n(z) \log \frac{\Gamma^n(z)}{\pi^0(z)} \geq \eta\right\} \\
&= \mathbb{I}\{D(\Gamma^n \|\pi^0) \geq \eta\} \\
&= \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta(\pi^0)\}
\end{aligned} \tag{4}$$

If we have some prior information on the alternate distribution π^1 , a different version of the GLRT is used. In particular, suppose it is known that the alternate distribution lies in a parametric family of distributions of the following form:

$$\mathcal{E}_{\pi^0} := \{\tilde{\pi}^r : r \in \mathbb{R}^d\}.$$

where $\tilde{\pi}^r \in \mathcal{P}(\mathcal{Z})$ are probability distributions on \mathcal{Z} parameterized by a parameter $r \in \mathbb{R}^d$. The specific form of $\tilde{\pi}^r$ is defined later in the paper. In this case, the resulting composite hypothesis testing problem is typically solved using a GLRT (see [3] for results related to the present paper, and [4] for a more recent account) of the following form:

$$\phi_n^{\text{MM}} = \mathbb{I}\left\{\sup_{\pi^1 \in \mathcal{E}_{\pi^0}} \sum_{z \in \mathcal{Z}} \Gamma^n(z) \log \frac{\pi^1(z)}{\pi^0(z)} \geq \eta\right\}. \tag{5}$$

We show that this test can be interpreted as a relaxation of the Hoeffding test of (4). In particular we show that (5) can be expressed in a form similar to (4),

$$\phi_n^{\text{MM}} = \mathbb{I}\{D^{\text{MM}}(\Gamma^n \|\pi^0) \geq \eta\} \tag{6}$$

where D^{MM} is the *mismatched divergence*; a relaxation of the K-L divergence, in the sense that $D^{\text{MM}}(\mu \|\pi) \leq D(\mu \|\pi)$ for any $\mu, \pi \in \mathcal{P}(\mathcal{Z})$. We refer to the test (6) as the *mismatched test*.

This paper is devoted to the analysis of the mismatched divergence and mismatched test.

The terminology is borrowed from the *mismatched channel* (see Lapidot [5] for a bibliography). The mismatched divergence described here is a generalization of the relaxation introduced in [6]. In this way we embed the analysis of the resulting universal test within the framework of Csiszár and Shields [7]. The mismatched test statistic can also be viewed as a generalization of the robust hypothesis testing statistic introduced in [8], [9].

When the alternate distribution satisfies $\pi^1 \in \mathcal{E}_{\pi^0}$, we show that, under some regularity conditions on \mathcal{E}_{π^0} , the mismatched test of (6) and Hoeffding's test of (4) have identical asymptotic performance in

terms of error exponents. A consequence of this result is that the GLRT is optimal in differentiating a particular distribution from others in an exponential family of distributions. We also establish that the proposed mismatched test has a significant advantage over the Hoeffding test in terms of finite sample size performance. This advantage is due to the difference in the asymptotic variances of the two test statistics under the null hypothesis. In particular, we show that the variance of the K-L divergence grows linearly with the alphabet size, making the test impractical for applications involving large alphabet distributions. We also show that the variance of the mismatched divergence grows linearly with the dimension d of the parameter space, and can hence be controlled through a prudent choice of the function class defining the mismatched divergence.

The remainder of the paper is organized as follows. We begin in Section II with a description of mismatched divergence and the mismatched test, and describe their relation to other concepts including robust hypothesis testing, composite hypothesis testing, reverse I-projection, and maximum likelihood (ML) estimation. Formulae for the asymptotic mean and variance of the test statistics are presented in Section III. Section III also contains a discussion interpreting these asymptotic results in terms of the performance of the detection rule. Proofs of the main results are provided in the appendix. Conclusions and directions for future research are contained in Section IV.

II. MISMATCHED DIVERGENCE

We adopt the following compact notation in the paper: For any function $f: \mathcal{Z} \rightarrow \mathbb{R}$ and $\pi \in \mathcal{P}(\mathcal{Z})$ we denote the mean $\sum_{z \in \mathcal{Z}} f(z)\pi(z)$ by $\pi(f)$, or by $\langle \pi, f \rangle$ when we wish to emphasize the convex-analytic setting. At times we will extend these definitions to allow functions f taking values in a vector space. For $z \in \mathcal{Z}$ and $\pi \in \mathcal{P}(\mathcal{Z})$, we still use $\pi(z)$ to denote the probability assigned to element z under measure π . The meaning of such notation will be clear from context.

The logarithmic moment generating function (log-MGF) is denoted

$$\Lambda_\pi(f) = \log(\pi(\exp(f)))$$

where $\pi(\exp(f)) = \sum_{z \in \mathcal{Z}} \pi(z) \exp(f(z))$ by the notation we introduced in the previous paragraph. For any two probability measures $\nu^1, \nu^2 \in \mathcal{P}(\mathcal{Z})$ the relative entropy is expressed,

$$D(\nu^1 \parallel \nu^2) = \begin{cases} \langle \nu^1, \log(\nu^1/\nu^2) \rangle & \text{if } \nu^1 \prec \nu^2 \\ \infty & \text{else} \end{cases}$$

where $\nu^1 \prec \nu^2$ denotes absolute continuity. The following proposition recalls a well-known variational representation. This can be obtained, for instance, by specializing the representation in [10] to an i.i.d. setting. An alternate variational representation of the divergence is introduced in [11].

Proposition II.1. *The relative entropy can be expressed as the convex dual of the log moment generating function: For any two probability measures $\nu^1, \nu^2 \in \mathcal{P}(\mathcal{Z})$,*

$$D(\nu^1 \parallel \nu^2) = \sup_f (\nu^1(f) - \Lambda_{\nu^2}(f)) \quad (7)$$

where the supremum is taken over the space of all real-valued functions on \mathcal{Z} . Furthermore, if ν^1 and ν^2 have equal supports, then the supremum is achieved by the log likelihood ratio function $f^* = \log(\nu^1/\nu^2)$.

Outline of proof: Although the result is well known, we provide a simple proof here since similar arguments will be reused later in the paper.

For any function f and probability measure ν we have,

$$\begin{aligned} D(\nu^1 \parallel \nu^2) &= \langle \nu^1, \log(\nu^1/\nu^2) \rangle \\ &= \langle \nu^1, \log(\nu/\nu^2) \rangle + \langle \nu^1, \log(\nu^1/\nu) \rangle \end{aligned}$$

On setting $\nu = \nu^2 \exp(f - \Lambda_{\nu^2}(f))$ this gives,

$$D(\nu^1 \parallel \nu^2) = \nu^1(f) - \Lambda_{\nu^2}(f) + D(\nu^1 \parallel \nu) \geq \nu^1(f) - \Lambda_{\nu^2}(f).$$

If ν^1 and ν^2 have equal supports, then the above inequality holds with equality for $f = \log(\nu^1/\nu^2)$, which would lead to $\nu = \nu^1$. This proves that (7) holds whenever ν^1 and ν^2 have equal supports. The proof for general distributions is similar and is omitted here. ■

The representation (7) is the basis of the mismatched divergence. We fix a set of functions denoted by \mathcal{F} , and obtain a lower bound on the relative entropy by taking the supremum over the smaller set as follows,

$$D^{\text{MM}}(\nu^1 \parallel \nu^2) := \sup_{f \in \mathcal{F}} \{ \nu^1(f) - \Lambda_{\nu^2}(f) \}. \quad (8)$$

If ν^1 and ν^2 have full support, and if the function class \mathcal{F} contains the log-likelihood ratio function $f^* = \log(\nu^1/\nu^2)$, then it is immediate from Proposition II.1 that the supremum in (8) is achieved by f^* , and in this case $D^{\text{MM}}(\nu^1 \parallel \nu^2) = D(\nu^1 \parallel \nu^2)$. Moreover, since the objective function in (8) is invariant to shifts of f , it follows that even if a constant scalar is added to the function f^* , it still achieves the supremum in (8).

In this paper the function class is assumed to be defined through a finite-dimensional parametrization of the form,

$$\mathcal{F} = \{f_r : r \in \mathbb{R}^d\} \quad (9)$$

Further assumptions will be imposed in our main results. In particular, we will assume that $f_r(z)$ is differentiable as a function of r for each z .

We fix a distribution $\pi \in \mathcal{P}(\mathcal{Z})$ and a function class of the form (9). For each $r \in \mathbb{R}^d$ the *twisted distribution* $\tilde{\pi}^r \in \mathcal{P}(\mathcal{Z})$ is defined as,

$$\tilde{\pi}^r := \pi \exp(f_r - \Lambda_\pi(f_r)). \quad (10)$$

The collection of all such distributions parameterized by r is denoted

$$\mathcal{E}_\pi := \{\tilde{\pi}^r : r \in \mathbb{R}^d\}. \quad (11)$$

A. Applications

The applications of mismatched divergence include those applications surveyed in Section 3 of [4] in their treatment of generalized likelihood ratio tests. Here we list potential applications in three domains: Hypothesis testing, source coding, and nonlinear filtering. Other applications include channel coding and signal detection, following [4].

1) *Hypothesis testing*: The problem of universal hypothesis testing is relevant in several practical applications including anomaly detection. It is often possible to have an accurate model of the normal behavior of a system, which is usually represented by the null hypothesis distribution π^0 . The anomalous behavior is often unknown, which is represented by the unknown alternate distribution. The primary motivation for our research is to improve the finite sample size performance of Hoeffding's universal hypothesis test (3). The difficulty we address is the large variance of this test statistic when the alphabet size is large. Theorem II.2 makes this precise:

Theorem II.2. *Let $\pi^0, \pi^1 \in \mathcal{P}(\mathcal{Z})$ have full supports over \mathcal{Z} .*

- (i) *Suppose that the observation sequence \mathbf{Z} is i.i.d. with marginal π^0 . Then the normalized Hoeffding test statistic sequence $\{nD(\Gamma^n \parallel \pi^0) : n \geq 1\}$ has the following asymptotic bias and variance:*

$$\lim_{n \rightarrow \infty} \mathbb{E}[nD(\Gamma^n \parallel \pi^0)] = \frac{1}{2}(N-1) \quad (12)$$

$$\lim_{n \rightarrow \infty} \text{Var}[nD(\Gamma^n \parallel \pi^0)] = \frac{1}{2}(N-1) \quad (13)$$

where $N = |\mathbf{Z}|$ denotes the size (cardinality) of \mathbf{Z} . Furthermore, the following weak convergence result holds:

$$nD(\Gamma^n \parallel \pi^0) \xrightarrow[n \rightarrow \infty]{d.} \frac{1}{2} \chi_{N-1}^2 \quad (14)$$

where the right hand side denotes the chi-squared distribution with $N - 1$ degrees of freedom.

(ii) Suppose the sequence \mathbf{Z} is drawn i.i.d. under $\pi^1 \neq \pi^0$. We then have,

$$\lim_{n \rightarrow \infty} \mathbb{E}[n(D(\Gamma^n \parallel \pi^0) - D(\pi^1 \parallel \pi^0))] = \frac{1}{2}(N - 1)$$

□

The bias result of (12) follows from the unpublished report [12] (see [13, Sec III.C]), and the weak convergence result of (14) is given in [14]. All the results of the theorem, including (13) also follow from Theorem III.2 — We elaborate on this in section III.

We see from Theorem II.2 that the bias of the divergence statistic $D(\Gamma^n \parallel \pi^0)$ decays as $\frac{N-1}{2n}$, irrespective of whether the observations are drawn from distribution π^0 or π^1 . One could argue that the problem of high bias in the Hoeffding test statistic can be addressed by setting a higher threshold. However, we also notice that when the observations are drawn under π^0 , the variance of the divergence statistic decays as $\frac{N-1}{2n^2}$, which can be significant when N is of the order of n^2 . This is a more serious flaw of the Hoeffding test for large alphabet sizes, since it cannot be addressed as easily.

The weak convergence result in (14), and other such results established later in this paper, can be used to guide the choice of thresholds for a finite sample test, subject to a constraint on the probability of false alarm (see for example, [7, p. 457]). As an application of (12) we propose the following approximation for the false alarm probability in the Hoeffding test defined in (4),

$$p_{\text{FA}} := \mathbb{P}_{\pi^0} \{ \phi_n^{\text{H}} = 1 \} \approx \mathbb{P} \left\{ \frac{1}{2} \sum_{i=1}^{N-1} W_i^2 \geq n\eta \right\} \quad (15)$$

where $\{W_i\}$ are i.i.d. $N(0, 1)$ random variables. In this way we can obtain a simple formula for the threshold to approximately achieve a given constraint on p_{FA} . For moderate values of the sequence length n , the χ^2 approximation gives a more accurate prediction of the false alarm probabilities for the Hoeffding test compared to those predicted using Sanov's theorem as we demonstrate below.

Consider the application of (15) in the following example. We used Monte-Carlo simulations to approximate the performance of the Hoeffding test described in (4), with π^0 the uniform distribution on an alphabet of size 20. Shown in Figure 1 is a semi-log plot comparing three quantities: The probability of false alarm p_{FA} , estimated via simulation; the approximation (15) obtained from the Central Limit

Theorem; and the approximation obtained from Sanov's Theorem, $\log(p_{\text{FA}}) \approx -n\eta$. It is clearly seen that the approximation based on the weak convergence result of (15) is *far more accurate* than the approximation based on Sanov's theorem. It should be noted that the approximate formula for the false alarm probability obtained from Sanov's theorem can be made more accurate by using refinements of large deviation results given in [15]. However, these refinements are often difficult to compute. For instance, it can be shown using the results of [15] that $p_{\text{FA}} \approx cn^{\frac{N-3}{2}} \exp(-n\eta)$ where constant c is given by a surface integral over the surface of the divergence ball, $\mathcal{Q}_\eta(\pi^0)$.

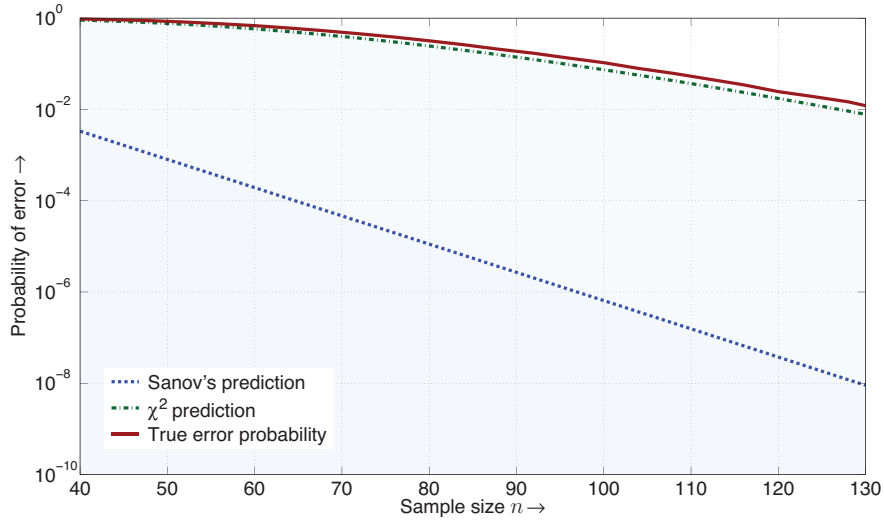


Fig. 1. Approximations for the error probability in universal hypothesis testing. The error probability of the Hoeffding test is closely approximated by the approximation (15).

One approach to addressing the implementation issues of the universal test is through clustering (or partitioning) the alphabet as in [16], or smoothing in the space of probability measures as in [17], [18] to extend the Hoeffding test to the case of continuous alphabets. The mismatched test proposed here is a generalization of a partition in the following sense. Suppose that $\{A_i : 1 \leq i \leq N_a\}$ are disjoint sets satisfying $\cup A_i = \mathcal{X}$, and let $Y(t) = i$ if $X(t) \in A_i$. Applying (13), we conclude that the Hoeffding test using \mathbf{Y} instead of \mathbf{X} will have asymptotic variance equal to $\frac{1}{2}(N_a - 1)$, where $N_a < N$ for a non-trivial partition. We have:

Proposition II.3. *Suppose that the mismatched divergence is defined with respect to the linear function class (26) using $\psi_i = \mathbb{I}_{A_i}$, $1 \leq i \leq N_a$. In this case the mismatched test (5) coincides with the Hoeffding test using observations \mathbf{Y} .* \square

The advantage of the mismatched test (5) over a partition is that we can incorporate prior knowledge regarding alternate statistics, and we can include non-standard ‘priors’ such as continuity of the log-likelihood ratio function between the null and alternate distributions. This is useful in anomaly detection applications where one may have models of anomalous behavior which can be used to design the correct mismatched test for the desired application.

2) *Source coding with training*: Let π denote a source distribution on a finite alphabet Z . Suppose we do not know π exactly and we design optimal codelengths assuming that the distribution is μ : For letter $z \in Z$ we let $\ell(z) = -\log(\mu(z))$ denote Shannon’s codeword length. The expected codelength is thus,

$$\mathbb{E}[\ell] = \sum_{z \in Z} \ell(z) \pi(z) = H(\pi) + D(\pi \| \mu)$$

where H denotes the entropy, $-\sum_{z \in Z} \pi(z) \log(\pi(z))$. Let $\ell^* := H(\pi)$ denote the optimal (minimal) expected codelength.

Now suppose it is known that under π the probability of each letter $z \in Z$ is bounded away from zero. That is, we assume that for some $\epsilon > 0$,

$$\pi \in \mathbb{P}_\epsilon := \{\mu \in \mathcal{P}(Z) : \mu(z) > \epsilon, \text{ for all } z \in Z\}.$$

Further suppose that a training sequence of length n is given, drawn under π . We are interested in constructing a source code for encoding symbols from the source π based on these training symbols. Let Γ^n denote the empirical distribution (i.e., the *type*) of the observations based on these n training symbols. We assign codeword lengths to each symbol z according to the following rule,

$$\ell(z) = \begin{cases} \log \frac{1}{\Gamma^n(z)} & \text{if } \Gamma^n \in \mathbb{P}_{\epsilon/2} \\ \log \frac{1}{\pi^u(z)} & \text{else} \end{cases}$$

where π^u is the uniform distribution on Z .

Let \mathcal{T} denote the sigma-algebra generated by the training symbols. The conditional expected codelength given \mathcal{T} satisfies,

$$\mathbb{E}[\ell^n | \mathcal{T}] = \begin{cases} \ell^* + D(\pi \| \Gamma^n) & \text{if } \Gamma^n \in \mathbb{P}_{\epsilon/2} \\ \ell^* + D(\pi \| \pi^u) & \text{else} \end{cases}$$

We study the behavior of $\mathbb{E}[\ell^n - \ell^* | \mathcal{T}]$ as a function of n . We argue in the appendix that a modification of the results from Theorem III.2 can be used to establish the following relations:

$$\begin{aligned} n(\mathbb{E}[\ell^n | \mathcal{T}] - \ell^*) &\xrightarrow[n \rightarrow \infty]{d.} \frac{1}{2} \chi_{N-1}^2 \\ \mathbb{E}[n(\ell^n - \ell^*)] &\xrightarrow[n \rightarrow \infty]{} \frac{1}{2}(N-1) \\ \text{Var}[n\mathbb{E}[\ell^n | \mathcal{T}]] &\xrightarrow[n \rightarrow \infty]{} \frac{1}{2}(N-1) \end{aligned} \tag{16}$$

where N is the cardinality of the alphabet \mathcal{Z} . Comparing with Theorem II.2 we conclude that the asymptotic behavior of the excess codelength is identical to the asymptotic behavior of the Hoeffding test statistic $D(\Gamma^n \parallel \pi)$ under π . Methods such as those proposed in this paper can be used to reduce high variance, just as in the hypothesis testing problem emphasized in this paper.

3) *Filtering*: The recent paper [19] considers approximations for the nonlinear filtering problem. Suppose that \mathbf{X} is a Markov chain on \mathbb{R}^n , and \mathbf{Y} is an associated observation process on \mathbb{R}^p of the form $Y(t) = \gamma(X(t), W(t))$, where \mathbf{W} is an i.i.d. sequence. The conditional distribution of $X(t)$ given $\{Y(0), \dots, Y(t)\}$ is denoted B_t ; known as the *belief state* in this literature. The evolution of the belief state can be expressed in a recursive form: For some mapping $\phi: \mathcal{B}(\mathbb{R}^n) \times \mathbb{R}^p \rightarrow \mathcal{B}(\mathbb{R}^n)$,

$$B_{t+1} = \phi(B_t, Y_{t+1}), \quad t \geq 0$$

The approximation proposed in [19] is based on a projection of B_t onto an exponential family of densities over \mathbb{R}^n , of the form $p_\theta(x) = p_0(x) \exp(\theta^T \psi(x) - \Lambda(\theta))$, $\theta \in \mathbb{R}^d$. They consider the *reverse I-projection*,

$$B^e = \arg \min_{\mu \in \mathcal{E}} D(B \parallel \mu)$$

where the minimum is over $\mathcal{E} = \{p_\theta\}$. From the definition of divergence this is equivalently expressed,

$$B^e = \arg \max_{\theta} \int \left(\theta^T \psi(x) - \Lambda(\theta) \right) B(dx) \quad (17)$$

A projected filter is defined by the recursion,

$$\hat{B}_{t+1} = [\phi(\hat{B}_t, Y_{t+1})]^e, \quad t \geq 0 \quad (18)$$

The techniques in the current paper provide algorithms for computation of this projection, and suggest alternative projection schemes, such as the robust approach described in Section II-F.

B. Basic structure of mismatched divergence

The mismatched test is defined to be a relaxation of the Hoeffding test described in (4). We replace the divergence functional with the mismatched divergence $D^{\text{MM}}(\Gamma^n \parallel \pi^0)$ to obtain the mismatched test sequence,

$$\phi_n^{\text{MM}} = \mathbb{I}\{D^{\text{MM}}(\Gamma^n \parallel \pi^0) \geq \eta\} = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta^{\text{MM}}(\pi^0)\} \quad (19)$$

where $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ is the mismatched divergence ball of radius η around π^0 defined analogously to (1):

$$\mathcal{Q}_\eta^{\text{MM}}(\mu) = \{\nu \in \mathcal{P}(\mathcal{Z}) : D^{\text{MM}}(\nu \parallel \mu) < \eta\}. \quad (20)$$

The next proposition establishes some basic geometry of the mismatched divergence balls. For any function g we define the following hyperplane and half-space:

$$\begin{aligned}\mathcal{H}_g &:= \{\nu : \nu(g) = 0\} \\ \mathcal{H}_g^- &:= \{\nu : \nu(g) < 0\}.\end{aligned}\tag{21}$$

Proposition II.4. *The following hold for any $\nu, \pi \in \mathcal{P}(\mathcal{Z})$, and any collection of functions \mathcal{F} :*

- (i) *For each $\eta > 0$ we have $\mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \bigcap \mathcal{H}_g^-$, where the intersection is over all functions g of the form,*

$$g = f - \Lambda_\pi(f) - \eta \tag{22}$$

with $f \in \mathcal{F}$.

- (ii) *Suppose that $\eta = D^{\text{MM}}(\nu \parallel \pi)$ is finite and non-zero. Further suppose that for $\nu^1 = \nu$ and $\nu^2 = \pi$, the supremum in (8) is achieved by $f^* \in \mathcal{F}$. Then \mathcal{H}_{g^*} is a supporting hyperplane to $\mathcal{Q}_\eta^{\text{MM}}(\pi)$, where g^* is given in (22) with $f = f^*$.*

Proof: (i) Suppose $\mu \in \mathcal{Q}_\eta^{\text{MM}}(\pi)$. Then, for any $f \in \mathcal{F}$,

$$\mu(f) - \Lambda_\pi(f) - \eta \leq D^{\text{MM}}(\mu \parallel \pi) - \eta < 0$$

That is, for any $f \in \mathcal{F}$, on defining g by (22) we obtain the desired inclusion $\mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \mathcal{H}_g^-$.

- (ii) Let $\mu \in \mathcal{H}_{g^*}$ be arbitrary. Then we have:

$$\begin{aligned}D^{\text{MM}}(\mu \parallel \pi) &= \sup_r (\mu(f_r) - \Lambda_\pi(f_r)) \\ &\geq \mu(f^*) - \Lambda_\pi(f^*) \\ &= \Lambda_\pi(f^*) + \eta - \Lambda_\pi(f^*) = \eta.\end{aligned}$$

Hence it follows that \mathcal{H}_{g^*} supports $\mathcal{Q}_\eta^{\text{MM}}(\pi)$ at ν . ■

C. Asymptotic optimality of the mismatched test

The asymptotic performance of a binary hypothesis testing problem is typically characterized in terms of error exponents. We adopt the following criterion for performance evaluation, following Hoeffding [2] (and others, notably [17], [18].) Suppose that the observations $\mathbf{Z} = \{Z_t : t = 1, \dots\}$ form an i.i.d. sequence evolving on \mathcal{Z} . For a given π^0 , and a given alternate distribution π^1 , the type I and type II error

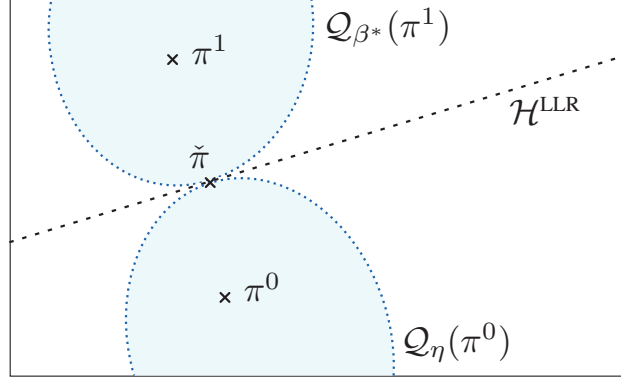


Fig. 2. Geometric interpretation of the log likelihood ratio test. The exponent $\beta^* = \beta^*(\eta)$ is the largest constant satisfying $Q_\eta(\pi^0) \cap Q_{\beta^*}(\pi^1) = \emptyset$. The hyperplane $\mathcal{H}^{\text{LLR}} := \{\nu : \nu(L) = \tilde{\pi}(L)\}$ separates the convex sets $Q_\eta(\pi^0)$ and $Q_{\beta^*}(\pi^1)$.

exponents are denoted respectively by,

$$\begin{aligned} J_\phi^0 &:= \liminf_{n \rightarrow \infty} -\frac{1}{n} \log(\mathbb{P}_{\pi^0}\{\phi_n = 1\}), \\ J_\phi^1 &:= \liminf_{n \rightarrow \infty} -\frac{1}{n} \log(\mathbb{P}_{\pi^1}\{\phi_n = 0\}) \end{aligned} \quad (23)$$

where in the first limit the marginal distribution of Z_t is π^0 , and in the second it is π^1 . The limit J_ϕ^0 is also called the false-alarm error exponent, and J_ϕ^1 the missed-detection error exponent.

For a given constraint $\eta > 0$ on the false-alarm exponent J_ϕ^0 , an optimal test is the solution to the asymptotic Neyman-Pearson hypothesis testing problem,

$$\beta^*(\eta) = \sup\{J_\phi^1 : \text{subject to } J_\phi^0 \geq \eta\} \quad (24)$$

where the supremum is over all allowed test sequences ϕ . While the exponent $\beta^*(\eta) = \beta^*(\eta, \pi^1)$ depends upon π^1 , Hoeffding's test we described in (4) does not require knowledge of π^1 , yet achieves the optimal exponent $\beta^*(\eta, \pi^1)$ for any π^1 . The optimality of Hoeffding's test established in [2] easily follows from Sanov's theorem.

While the mismatched test described in (6) is not always optimal for (24) for a general choice of π^1 , it is optimal for some specific choices of the alternate distributions. The following corollary to Proposition II.4 captures this idea.

Corollary II.1. *Suppose $\pi^0, \pi^1 \in \mathcal{P}(\mathcal{Z})$ have equal supports. Further suppose that for all $\alpha > 0$, there exists $\tau \in \mathbb{R}$ and $r \in \mathbb{R}^d$ such that*

$$\alpha L(z) + \tau = f_r(z) \quad \text{a.e.} \quad [\pi^0],$$

where L is the log likelihood-ratio function $L := \log(\pi^1/\pi^0)$. Then the mismatched test is optimal in the sense that the constraint $J_{\phi^{\text{MM}}}^0 \geq \eta$ is satisfied with equality, and under π^1 the optimal error exponent is achieved; i.e. $J_{\phi^{\text{MM}}}^1 = \beta^*(\eta)$ for all $\eta \in (0, D(\pi^1\|\pi^0))$.

Proof: Suppose that the conditions stated in the corollary hold. Consider the twisted distribution $\tilde{\pi} = \kappa(\pi^0)^{1-\varrho}(\pi^1)^\varrho$, where κ is a normalizing constant and $\varrho \in (0, 1)$ is chosen so as to guarantee $D(\tilde{\pi}\|\pi^0) = \eta$. It is known that the hyperplane $\mathcal{H}^{\text{LLR}} := \{\nu : \nu(L) = \tilde{\pi}(L)\}$ separates the divergence balls $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ at $\tilde{\pi}$. This geometry, which is implicit in [17], is illustrated in Figure 2.

From the form of $\tilde{\pi}$ it is also clear that

$$\log \frac{\tilde{\pi}}{\pi^0} = \varrho L - \Lambda_{\pi^0}(\varrho L).$$

Hence it follows that the supremum in the variational representation of $D(\tilde{\pi}\|\pi^0)$ is achieved by ϱL . Furthermore, since $\varrho L + \tau \in \mathcal{F}$ for some $\tau \in \mathbb{R}$ we have

$$\begin{aligned} D^{\text{MM}}(\tilde{\pi}\|\pi^0) &= D(\tilde{\pi}\|\pi^0) = \eta \\ &= \tilde{\pi}(\varrho L + \tau) - \Lambda_{\pi^0}(\varrho L + \tau) \\ &= \tilde{\pi}(\varrho L) - \Lambda_{\pi^0}(\varrho L). \end{aligned}$$

This means that $\mathcal{H}^{\text{LLR}} = \{\nu : \nu(\varrho L - \Lambda_{\pi^0}(\varrho L) - \eta) = 0\}$. Hence, by applying Proposition II.4 (ii) it follows that the hyperplane \mathcal{H}^{LLR} separates $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$. This in particular means that the sets $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ are disjoint. This fact, together with Sanov's theorem proves the corollary. ■

The corollary indicates that while using the mismatched test in practice, the function class might be chosen to include approximations to scaled versions of the log-likelihood ratio functions of the anticipated alternate distributions $\{\pi^1\}$ with respect to π^0 .

The mismatched divergence has several equivalent characterizations. We first relate it to an ML estimate from a parametric family of distributions.

D. Mismatched divergence and ML estimation

On interpreting $f_r - \Lambda_\pi(f_r)$ as a log-likelihood ratio function we obtain in Proposition II.5 the following representation of mismatched divergence,

$$D^{\text{MM}}(\mu\|\pi) = \sup_{r \in \mathbb{R}^d} (\mu(f_r) - \Lambda_\pi(f_r)) = D(\mu\|\pi) - \inf_{\nu \in \mathcal{E}_\pi} D(\mu\|\nu). \quad (25)$$

The infimum on the RHS of (25) is known as *reverse I-projection* [7]. Proposition II.6 that follows uses this representation to obtain other interpretations of the mismatched test.

Proposition II.5. *The identity (25) holds for any function class \mathcal{F} . The supremum is achieved by some $r^* \in \mathbb{R}^d$ if and only if the infimum is attained at $\nu^* = \tilde{\pi}^{r^*} \in \mathcal{E}_\pi$. If a minimizer ν^* exists, we obtain the generalized Pythagorean identity,*

$$D(\mu\|\pi) = D^{\text{MM}}(\mu\|\pi) + D(\mu\|\nu^*)$$

Proof: For any r we have $\mu(f_r) - \Lambda_\pi(f_r) = \mu(\log(\tilde{\pi}^r/\pi))$. Consequently,

$$\begin{aligned} D^{\text{MM}}(\mu\|\pi) &= \sup_r (\mu(f_r) - \Lambda_\pi(f_r)) \\ &= \sup_r \mu \left(\log \left(\frac{\mu \tilde{\pi}^r}{\pi \mu} \right) \right) \\ &= \sup_r \{D(\mu\|\pi) - D(\mu\|\tilde{\pi}^r)\} \end{aligned}$$

This proves the identity (25), and the remaining conclusions follow directly. ■

The representation of Proposition II.5 invites the interpretation of the optimizer in the definition of the mismatched test statistic in terms of an ML estimate. Given the well-known correspondence between maximum-likelihood estimation and the generalized likelihood ratio test (GLRT), Proposition II.6 implies that the mismatched test is a special case of the GLRT analyzed in [3].

Proposition II.6. *Suppose that the observations Z are modeled as an i.i.d. sequence, with marginal in the family \mathcal{E}_π . Let \hat{r}^n denote the ML estimate of r based on the first n samples,*

$$\begin{aligned} \hat{r}^n &\in \arg \max_{r \in \mathbb{R}^d} \mathbb{P}_{\tilde{\pi}^r} \{Z_1 = a_1, Z_2 = a_2, \dots, Z_n = a_n\} \\ &= \arg \max_{r \in \mathbb{R}^d} \prod_{i=1}^n \tilde{\pi}^r(a_i) \end{aligned}$$

where a_i indicates the observed value of the i -th symbol. Assuming the maximum is attained we have the following interpretations:

(i) *The distribution $\tilde{\pi}^{\hat{r}^n}$ solves the reverse I-projection problem,*

$$\tilde{\pi}^{\hat{r}^n} \in \arg \min_{\nu \in \mathcal{E}_\pi} D(\Gamma^n\|\nu).$$

(ii) *The function $f^* = f_{\hat{r}^n}$ achieves the supremum that defines the mismatched divergence, $D^{\text{MM}}(\Gamma^n\|\pi) = \Gamma^n(f^*) - \Lambda_\pi(f^*)$.*

Proof: The ML estimate can be expressed $\hat{r}^n = \arg \max_{r \in \mathbb{R}^d} \langle \Gamma^n, \log \tilde{\pi}^r \rangle$, and hence (i) follows by the identity,

$$\arg \min_{\nu \in \mathcal{E}_\pi} D(\Gamma^n\|\nu) = \arg \max_{\nu \in \mathcal{E}_\pi} \langle \Gamma^n, \log \nu \rangle, \quad \nu \in \mathcal{P}.$$

Combining the result of part (i) with Proposition II.5 we get the result of part (ii). ■

From conclusions of Proposition II.5 and Proposition II.6 we have,

$$\begin{aligned} D^{\text{MM}}(\Gamma^n \parallel \pi) &= \langle \Gamma^n, \log \frac{\tilde{\pi}^{\hat{f}^n}}{\pi} \rangle \\ &= \max_{\nu \in \mathcal{E}_\pi} \langle \Gamma^n, \log \frac{\nu}{\pi} \rangle \\ &= \max_{\nu \in \mathcal{E}_\pi} \frac{1}{n} \sum_{i=1}^n \log \frac{\nu(Z_i)}{\pi(Z_i)}. \end{aligned}$$

In general when the supremum in the definition of $D^{\text{MM}}(\Gamma^n \parallel \pi)$ may not be achieved, the maxima in the above equations are replaced with suprema and we have the following identity:

$$D^{\text{MM}}(\Gamma^n \parallel \pi) = \sup_{\nu \in \mathcal{E}_\pi} \frac{1}{n} \sum_{i=1}^n \log \frac{\nu(Z_i)}{\pi(Z_i)}.$$

Thus the test statistic used in the mismatched test of (6) is exactly the generalized likelihood ratio between the family of distributions \mathcal{E}_{π^0} and π^0 where

$$\mathcal{E}_{\pi^0} = \{\pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)) : r \in \mathbb{R}^d\}.$$

More structure can be established when the function class is linear.

E. Linear function class and I-projection

The mismatched divergence introduced in [6] was restricted to a linear function class. Let $\{\psi_i : 1 \leq i \leq d\}$ denote d functions on \mathcal{Z} , let $\psi = (\psi_1, \dots, \psi_d)^T$, and let $f_r = r^T \psi$ in the definition (9):

$$\mathcal{F} = \left\{ f_r = \sum_{i=1}^d r_i \psi_i : r \in \mathbb{R}^d \right\}. \quad (26)$$

A linear function class is particularly appealing because the optimization problem in (8) used to define the mismatched divergence becomes a convex program and hence is easy to evaluate in practice. Furthermore, for such a linear function class, the collection of twisted distributions \mathcal{E}_π defined in (11) forms an *exponential family* of distributions.

Proposition II.5 expresses $D^{\text{MM}}(\mu \parallel \pi)$ as a difference between the ordinary divergence and the value of a reverse I-projection $\inf_{\nu \in \mathcal{E}_\pi} D(\mu \parallel \nu)$. The next result establishes a characterization in terms of a (forward) I-projection. For a given vector $c \in \mathbb{R}^d$ we let \mathbb{P} denote the *moment class*

$$\mathbb{P} = \{\nu \in \mathcal{P}(\mathcal{Z}) : \nu(\psi) = c\} \quad (27)$$

where $\nu(\psi) = (\nu(\psi_1), \nu(\psi_2), \dots, \nu(\psi_d))^T$.

Proposition II.7. Suppose that the supremum in the definition of $D^{\text{MM}}(\mu\|\pi)$ is achieved at some $r^* \in \mathbb{R}^d$.

Then,

(i) The distribution $\nu^* := \check{\pi}^{r^*} \in \mathcal{E}_\pi$ satisfies,

$$D^{\text{MM}}(\mu\|\pi) = D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\},$$

where \mathbb{P} is defined using $c = \mu(\psi)$ in (27).

(ii) $D^{\text{MM}}(\mu\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}$, where g^* is given in (22) with $f = r^{*T}\psi$, and $\eta = D^{\text{MM}}(\mu\|\pi)$.

Proof: Since the supremum is achieved, the gradient must vanish by the first order condition for optimality:

$$\nabla(\mu(f_r) - \Lambda_\pi(f_r)) \Big|_{r=r^*} = 0$$

The gradient is computable, and the identity above can thus be expressed $\mu(\psi) - \check{\pi}^{r^*}(\psi) = 0$. That is, the first order condition for optimality is equivalent to the constraint $\check{\pi}^{r^*} \in \mathbb{P}$. Consequently,

$$\begin{aligned} D(\nu^*\|\pi) &= \langle \check{\pi}^{r^*}, \log \frac{\check{\pi}^{r^*}}{\pi} \rangle \\ &= \check{\pi}^{r^*}(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) \\ &= \mu(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) = D^{\text{MM}}(\mu\|\pi). \end{aligned}$$

Furthermore, by the convexity of $\Lambda_\pi(f_r)$ in r , it follows that the optimal r^* in the definition of $D^{\text{MM}}(\nu\|\pi)$ is the same for all $\nu \in \mathbb{P}$. Hence, it follows by the Pythagorean equality of Proposition II.5 that

$$D(\nu\|\pi) = D(\nu\|\nu^*) + D(\nu^*\|\pi), \text{ for all } \nu \in \mathbb{P}.$$

Minimizing over $\nu \in \mathbb{P}$ it follows that ν^* is the I-projection of π onto \mathbb{P} :

$$D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\}$$

which gives (i).

To establish (ii), note first that by (i) and the inclusion $\mathbb{P} \subset \mathcal{H}_{g^*}$ we have,

$$\begin{aligned} D^{\text{MM}}(\mu\|\pi) &= \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\} \\ &\geq \inf\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}. \end{aligned}$$

The reverse inequality follows from Proposition II.4 (i), and moreover the infimum is achieved with ν^* . ■

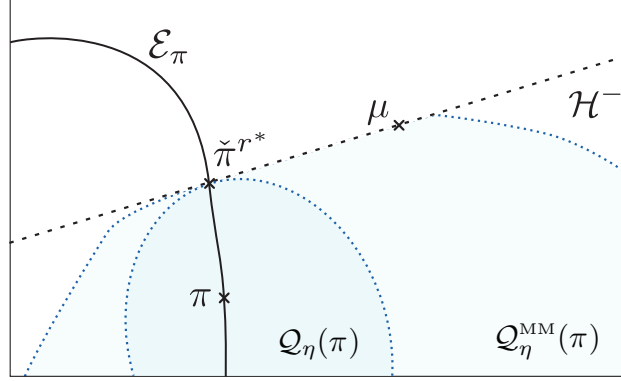


Fig. 3. *Interpretations of the mismatched divergence for a linear function class.* The distribution $\tilde{\pi}^{r^*}$ is the I-projection of π onto a hyperplane $\mathcal{H}_{g^*}^-$. It is also the reverse I-projection of μ onto the exponential family \mathcal{E}_π .

The geometry underlying mismatched divergence for a linear function class is illustrated in Figure 3. Suppose that the assumptions of Proposition II.7 hold, so that the supremum in (25) is achieved at r^* . Let $\eta = D^{\text{MM}}(\mu||\pi) = \mu(f_{r^*}) - \Lambda_\pi(f_{r^*})$, and $g^* = f_{r^*} - (\eta + \Lambda_\pi(f_{r^*}))$. Proposition II.4 implies that $\mathcal{H}_{g^*}^-$ defines a hyperplane passing through μ , with $\mathcal{Q}_\eta(\pi) \subset \mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \mathcal{H}_{g^*}^-$. This is strengthened in the linear case by Proposition II.7, which states that $\mathcal{H}_{g^*}^-$ supports $\mathcal{Q}_\eta(\pi)$ at the distribution $\tilde{\pi}^{r^*}$. Furthermore Proposition II.5 asserts that the distribution $\tilde{\pi}^{r^*}$ minimizes $D(\mu||\tilde{\pi})$ over all $\tilde{\pi} \in \mathcal{E}_\pi$.

The result established in Corollary II.1 along with the interpretation of the mismatched test as a GLRT can be used to show that the GLRT is asymptotically optimal for an exponential family of distributions.

Theorem II.8. *Let π^0 be some probability distribution over a finite set \mathcal{Z} . Let \mathcal{F} be a linear function class as defined in (26) and \mathcal{E}_{π^0} be the associated exponential family of distributions defined in (11). Consider the generalized likelihood ratio test (GLRT) between π^0 and \mathcal{E}_{π^0} defined by the following sequence of decision rules:*

$$\phi_n^{\text{GLRT}} = \mathbb{I}\left\{ \sup_{\nu \in \mathcal{E}_{\pi^0}} \frac{1}{n} \sum_{i=1}^n \log \frac{\nu(Z_i)}{\pi^0(Z_i)} \geq \eta \right\}.$$

The GLRT solves the composite hypothesis testing problem (24) for all $\pi^1 \in \mathcal{E}_{\pi^0}$ in the sense that the constraint $J_{\phi^{\text{GLRT}}}^0 \geq \eta$ is satisfied with equality, and under π^1 the optimal error exponent $\beta^(\eta)$ is achieved for all $\eta \in (0, D(\pi^1||\pi^0))$ and for all $\pi^1 \in \mathcal{E}_{\pi^0}$; i.e., $J_{\phi^{\text{GLRT}}}^1 = \beta^*(\eta)$.*

Proof: From Proposition II.6 and the discussion following the proposition, we know that ϕ^{GLRT} is the same as the mismatched test defined with respect to the function class \mathcal{F} . Moreover, any distribution $\pi^1 \in \mathcal{E}_{\pi^0}$ is of the form $\tilde{\pi}^r = \pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r))$ for some $r \in \mathbb{R}^d$ as defined in (10). Using L to

denote the log-likelihood ratio function between π^1 and π^0 , it follows by the linearity of \mathcal{F} that for any $\alpha > 0$,

$$\begin{aligned}\alpha L &= \alpha(f_r - \Lambda_{\pi^0}(f_r)) \\ &= f_{\alpha r} + \tau\end{aligned}$$

for some $\tau \in \mathbb{R}$. Hence, it follows by the conclusion of Corollary II.1 that the GLRT ϕ^{GLRT} solves the composite hypothesis testing problem (24) between π^0 and \mathcal{E}_{π^0} . ■

The above result is a special case of the sufficient conditions for optimality of the GLRT established in [3, Thm 2, p. 1600]. From the proof it is easily seen that the result can be extended to hold for composite hypothesis tests between π^0 and any family of distributions \mathcal{E}_{π^0} of the form in (11) provided \mathcal{F} is closed under positive scaling. It is also possible to strengthen the result of Corollary II.1 to obtain an alternate proof of [3, Thm 2, p. 1600]. We refer the reader to [20] for details.

F. Log-linear function class and robust hypothesis testing

In the prior work [8], [9] the following relaxation of entropy is considered,

$$D^{\text{ROB}}(\mu \parallel \pi) := \inf_{\nu \in \mathbb{P}} D(\mu \parallel \nu) \quad (28)$$

where the moment class \mathbb{P} is defined in (27) with $c = \pi(\psi)$, for a given collection of functions $\{\psi_i : 1 \leq i \leq d\}$. The associated universal test solves a min-max robust hypothesis testing problem.

We show here that D^{ROB} coincides with D^{MM} for a particular function class. It is described as (9) in which each function f_r is of the log-linear form,

$$f_r = \log(1 + r^T \psi)$$

subject to the constraint that $1 + r^T \psi(z)$ is strictly positive for each z . We further require that the functions ψ have zero mean under distribution π - i.e., we require $\pi(\psi) = 0$.

Proposition II.9. *For a given $\pi \in \mathcal{P}(\mathcal{Z})$, suppose that the log-linear function class \mathcal{F} is chosen with functions $\{\psi_i\}$ satisfying $\pi(\psi) = 0$. Suppose that the moment class used in the definition of D^{ROB} is chosen consistently, with $c = 0$ in (27). We then have for each $\mu \in \mathcal{P}(\mathcal{Z})$,*

$$D^{\text{MM}}(\mu \parallel \pi) = D^{\text{ROB}}(\mu \parallel \pi)$$

Proof: For each $\mu \in \mathcal{P}(\mathcal{Z})$, we obtain the following identity by applying Theorem 1.4 in [9],

$$\inf_{\nu \in \mathbb{P}} D(\mu \parallel \nu) = \sup \{ \mu(\log(1 + r^T \psi)) : 1 + r^T \psi(z) > 0 \text{ for all } z \in \mathcal{Z} \}$$

Moreover, under the assumption that $\pi(\psi) = 0$ we obtain,

$$\Lambda_\pi(\log(1 + r^T \psi)) = \log(\pi(1 + r^T \psi)) = 0$$

Combining these identities gives,

$$\begin{aligned} D^{\text{ROB}}(\mu \parallel \pi) &:= \inf_{\nu \in \mathbb{P}} D(\mu \parallel \nu) \\ &= \sup \{ \mu(\log(1 + r^T \psi)) - \Lambda_\pi(\log(1 + r^T \psi)) : \\ &\quad 1 + r^T \psi(z) > 0 \text{ for all } z \in Z \} \\ &= \sup_{f \in \mathcal{F}} \{ \mu(f) - \Lambda_\pi(f) \} = D^{\text{MM}}(\mu \parallel \pi) \end{aligned}$$

■

III. ASYMPTOTIC STATISTICS

In this section, we analyze the asymptotic statistics of the mismatched test. We require some assumptions regarding the function class $\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}$ to establish these results. Note that the second and third assumptions given below involve a distribution $\mu^0 \in \mathcal{P}(Z)$, and a vector $s \in \mathbb{R}^d$. We will make specialized versions of these assumptions in establishing our results, based on specific values of μ^0 and s . We use $Z_{\mu^0} \subset Z$ to denote the support of μ^0 and $\mathcal{P}(Z_{\mu^0})$ to denote the space of probability measures supported on Z_{μ^0} , viewed as a subset of $\mathcal{P}(Z)$.

Assumptions

- (A1) $f_r(z)$ is C^2 in r for each $z \in Z$.
- (A2) There exists a neighborhood B of μ^0 , open in $\mathcal{P}(Z_{\mu^0})$ such that for each $\mu \in B$, the supremum in the definition of $D^{\text{MM}}(\mu \parallel \mu^0)$ in (8) is achieved at a unique point $r(\mu)$.
- (A3) The vectors $\{\psi_0, \dots, \psi_d\}$ are linearly independent over the support of μ^0 , where $\psi_0 \equiv 1$, and for each $i \geq 1$

$$\psi_i(z) = \left. \frac{\partial}{\partial r_i} f_r(z) \right|_{r=s}, \quad z \in Z. \quad (29)$$

The linear-independence assumption in (A3) is defined as follows: If there are constants $\{a_0, \dots, a_d\}$ satisfying $\sum_{i=1}^d a_i \psi_i(z) = 0$ a.e. $[\mu^0]$, then $a_i = 0$ for each i . In the case of a linear function class, the functions $\{\psi_i, i \geq 1\}$ defined in (29) are just the basis functions in (26). Lemma III.1 provides an alternate characterization of Assumption (A3).

For any $\mu \in \mathcal{P}(Z)$ define the covariance matrix Σ_μ via,

$$\Sigma_\mu(i, j) = \mu(\psi_i \psi_j) - \mu(\psi_i) \mu(\psi_j), \quad 1 \leq i, j \leq d. \quad (30)$$

We use $\text{Cov}_\mu(g)$ to denote the covariance of an arbitrary real-valued function g under μ :

$$\text{Cov}_\mu(g) := \mu(g^2) - \mu(g)^2. \quad (31)$$

Lemma III.1. *Assumption (A3) holds if and only if $\Sigma_{\mu^0} > 0$.*

Proof: We evidently have $v^T \Sigma_{\mu^0} v = \text{Cov}_{\mu^0}(v^T \psi) \geq 0$ for any vector $v \in \mathbb{R}^d$. Hence, we have the following equivalence: For any $v \in \mathbb{R}^d$, on denoting $c_v = \mu^0(v^T \psi)$,

$$v^T \Sigma_{\mu^0} v = 0 \quad \Leftrightarrow \quad \sum_{i=1}^d v_i \psi_i(z) = c_v \quad \text{a.e. } [\mu^0]$$

The conclusion of the lemma follows. ■

We now present our main asymptotic results. Theorem III.2 identifies the asymptotic bias and variance of the mismatched test statistic under the null hypothesis, and also under the alternate hypothesis. A key observation is that the asymptotic bias and variance does not depend on N , the cardinality of \mathcal{Z} .

Theorem III.2. *Suppose that the observation sequence \mathbf{Z} is i.i.d. with marginal π . Suppose that there exists r^* satisfying $f_{r^*} = \log(\pi/\pi^0)$. Further, suppose that Assumptions (A1), (A2), (A3) hold with $\mu^0 = \pi$ and $s = r^*$. Then,*

(i) *When $\pi = \pi^0$,*

$$\lim_{n \rightarrow \infty} \mathbb{E}[nD^{\text{MM}}(\Gamma^n \parallel \pi^0)] = \frac{1}{2}d \quad (32)$$

$$\lim_{n \rightarrow \infty} \text{Var}[nD^{\text{MM}}(\Gamma^n \parallel \pi^0)] = \frac{1}{2}d \quad (33)$$

$$nD^{\text{MM}}(\Gamma^n \parallel \pi^0) \xrightarrow[n \rightarrow \infty]{d.} \frac{1}{2}\chi_d^2$$

(ii) *When $\pi = \pi^1 \neq \pi^0$, we have with $\sigma_1^2 := \text{Cov}_{\pi^1}(f_{r^*})$,*

$$\lim_{n \rightarrow \infty} \mathbb{E}[n(D^{\text{MM}}(\Gamma^n \parallel \pi^0) - D(\pi^1 \parallel \pi^0))] = \frac{1}{2}d \quad (34)$$

$$\lim_{n \rightarrow \infty} \text{Var}[n^{\frac{1}{2}}D^{\text{MM}}(\Gamma^n \parallel \pi^0)] = \sigma_1^2 \quad (35)$$

$$n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n \parallel \pi^0) - D(\pi^1 \parallel \pi^0)) \xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \sigma_1^2). \quad (36)$$

□

In part (ii) of Theorem III.2, the assumption that r^* exists implies that π^1 and π^0 have equal supports. Furthermore, if Assumption (A3) holds in part (ii), then a sufficient condition for Assumption (A2) is that the function $V(r) := (-\pi^1(f_r) + \Lambda_{\pi^0}(f_r))$ be coercive in r . And, under (A3), the function V is strictly

convex and coercive in the following settings: (i) If the function class is linear, or (ii) the function class is log-linear, and the two distributions π^1 and π^0 have common support. We use this fact in Theorem III.3 for the linear function class. The assumption of the existence of r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$ in part (ii) of Theorem III.2 can be relaxed. In the case of a linear function class we have the following extension of part (ii).

Theorem III.3. *Suppose that the observation sequence \mathbf{Z} is drawn i.i.d. with marginal π^1 satisfying $\pi^1 \prec \pi^0$. Let \mathcal{F} be the linear function class defined in (26). Suppose the supremum in the definition of $D^{\text{MM}}(\pi^1 \|\pi^0)$ is achieved at some $r^1 \in \mathbb{R}^d$. Further, suppose that the functions $\{\psi_i\}$ satisfy the linear independence condition of Assumption (A3) with $\mu^0 = \pi^1$. Then we have,*

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}[n(D^{\text{MM}}(\Gamma^n \|\pi^0) - D^{\text{MM}}(\pi^1 \|\pi^0))] &= \frac{1}{2} \text{trace}(\Sigma_{\pi^1} \Sigma_{\tilde{\pi}}^{-1}) \\ \lim_{n \rightarrow \infty} \text{Var}[n^{\frac{1}{2}} D^{\text{MM}}(\Gamma^n \|\pi^0)] &= \sigma_1^2 \\ n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n \|\pi^0) - D^{\text{MM}}(\pi^1 \|\pi^0)) &\xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \sigma_1^2) \end{aligned}$$

where in the first limit $\tilde{\pi} = \pi^0 \exp(f_{r^1} - \Lambda_{\pi^0}(f_{r^1}))$, and Σ_{π^1} and $\Sigma_{\tilde{\pi}}$ are defined as in (30). In the second two limits $\sigma_1^2 = \text{Cov}_{\pi^1}(f_{r^1})$. \square

Although we have not explicitly imposed Assumption (A2) in Theorem III.3, the argument we presented following Theorem III.2 ensures that when $\pi^1 \prec \pi^0$, Assumption (A2) is satisfied whenever Assumption (A3) holds. Furthermore, it can be shown that the achievement of the supremum required in Theorem III.3 is guaranteed if π^1 and π^0 have equal supports. We also note that the vector s appearing in eq. (29) of Assumption (A3) is arbitrary when the parametrization of the function class is linear.

The weak convergence results in Theorem III.2 (i) can be derived from Clarke and Barron [12], [13] (see also [7, Theorem 4.2]), following the maximum-likelihood estimation interpretation of the mismatched test obtained in Proposition II.6. In the statistics literature, such results are called *Wilks phenomena* after the initial work by Wilks [14].

These results can be used to set thresholds for a target false alarm probability in the mismatched test, just like we did for the Hoeffding test in (15). It is shown in [21] that such results can be used to set thresholds for the robust hypothesis testing problem described in Section II-F.

Implications for Hoeffding test The divergence can be interpreted as a special case of mismatched divergence defined with respect to a linear function class. Using this interpretation, the results of Theorem III.2 can also be specialized to obtain results on the Hoeffding test statistic. To satisfy the uniqueness condition

of Assumption (A2), we require that the function class should not contain any constant functions. Now suppose that the span of the linear function class \mathcal{F} together with the constant function $f^0 \equiv 1$ spans the set of all functions on \mathcal{Z} . This together with Assumption (A3) would imply that $d = N - 1$, where N is the size of the alphabet \mathcal{Z} . It follows from Proposition II.1 that for such a function class the mismatched divergence coincides with the divergence. Thus, an application of Theorem III.2 (i) gives rise to the results stated in Theorem II.2.

To prove Theorem III.2 and Theorem III.3 we need the following lemmas, whose proofs are given in the Appendix.

The following lemma will be used to deduce part (ii) of Theorem III.2 from part (i).

Lemma III.4. *Let $D_{\mathcal{F}}^{\text{MM}}$ denote the mismatched divergence defined using function class \mathcal{F} . Suppose $\pi^1 \prec \pi^0$ and the supremum in the definition of $D_{\mathcal{F}}^{\text{MM}}(\pi^1 \|\pi^0)$ is achieved at some $f_{r^*} \in \mathcal{F}$. Let $\tilde{\pi} = \pi^0 \exp(f_{r^*} - \Lambda_{\pi^0}(f_{r^*}))$ and $\mathcal{G} = \mathcal{F} - f_{r^*} := \{f_r - f_{r^*} : r \in \mathbb{R}^d\}$. Then for any μ satisfying $\mu \prec \pi^0$, we have*

$$D_{\mathcal{F}}^{\text{MM}}(\mu \|\pi^0) = D_{\mathcal{F}}^{\text{MM}}(\pi^1 \|\pi^0) + D_{\mathcal{G}}^{\text{MM}}(\mu \|\tilde{\pi}) + \langle \mu - \pi^1, \log(\frac{\tilde{\pi}}{\pi^0}) \rangle. \quad (37)$$

□

Suppose we apply the decomposition result from Lemma III.4 to the type of the observation sequence \mathbf{Z} , assumed to be drawn i.i.d. with marginal π^1 . If there exists r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$, then we have $\tilde{\pi} = \pi^1$. The decomposition becomes

$$D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \|\pi^0) = D_{\mathcal{F}}^{\text{MM}}(\pi^1 \|\pi^0) + D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \|\pi^1) + \langle \Gamma^n - \pi^1, f_{r^*} \rangle. \quad (38)$$

For large n , the second term in the decomposition (38) has a mean of order n^{-1} and variance of order n^{-2} , as shown in part (i) of Theorem III.2. The third term has zero mean and variance of order n^{-1} , since by the Central Limit Theorem,

$$n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle \xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \text{Cov}_{\pi^1}(f_{r^*})). \quad (39)$$

Thus, the asymptotic variance of $D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \|\pi^0)$ is dominated by that of the third term and the asymptotic bias is dominated by that of the second term. Thus we see that part (ii) of Theorem III.2 can be deduced from part (i).

Lemma III.5. *Let $\mathbf{X} = \{X^i : i = 1, 2, \dots\}$ be an i.i.d. sequence with mean \bar{x} taking values in a compact convex set $\mathcal{X} \subset \mathbb{R}^m$, containing \bar{x} as a relative interior point. Define $S^n = \frac{1}{n} \sum_{i=1}^n X^i$. Suppose we are*

given a function $h : \mathbb{R}^m \mapsto \mathbb{R}$, together with a compact set K containing \bar{x} as a relative interior point such that,

- 1) The gradient $\nabla h(x)$ and the Hessian $\nabla^2 h(x)$ are continuous over a neighborhood of K .
- 2) $\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\{S^n \notin K\} > 0$.

Let $M = \nabla^2 h(\bar{x})$ and $\Xi = \text{Cov}(X^1)$. Then,

- (i) The normalized asymptotic bias of $\{h(S^n) : n \geq 1\}$ is obtained via,

$$\lim_{n \rightarrow \infty} n\mathbb{E}[h(S^n) - h(\bar{x})] = \frac{1}{2}\text{trace}(M\Xi)$$

- (ii) If in addition to the above conditions, the directional derivative satisfies $\nabla h(\bar{x})^T(X^1 - \bar{x}) = 0$ almost surely, then the asymptotic variance decays as n^{-2} , with

$$\lim_{n \rightarrow \infty} n^2 \text{Var}[h(S^n)] = \frac{1}{2}\text{trace}(M\Xi M\Xi)$$

□

Lemma III.6. Suppose that the observation sequence \mathbf{Z} is drawn i.i.d. with marginal $\mu \in \mathcal{P}(\mathbf{Z})$. Let $h : \mathcal{P}(\mathbf{Z}) \mapsto \mathbb{R}$ be a continuous real-valued function whose gradient and Hessian are continuous in a neighborhood of μ . If the directional derivative satisfies $\nabla h(\mu)^T(\nu - \mu) \equiv 0$ for all $\nu \in \mathcal{P}(\mathbf{Z})$, then

$$n(h(\Gamma^n) - h(\mu)) \xrightarrow[n \rightarrow \infty]{d.} \frac{1}{2}W^T M W \quad (40)$$

where $M = \nabla^2 h(\mu)$ and $W \sim \mathcal{N}(0, \Sigma_W)$ with $\Sigma_W = \text{diag}(\mu) - \mu\mu^T$. □

Lemma III.7. Suppose that V is an m -dimensional, $\mathcal{N}(0, I_m)$ random variable, and $D : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a projection matrix. Then $\xi := \|DV\|^2$ is a chi-squared random variable with K degrees of freedom, where K denotes the rank of D . □

Before we proceed to the proofs of Theorem III.2 and Theorem III.3, we recall the optimization problem (25) defining the mismatched divergence:

$$D^{\text{MM}}(\mu \parallel \pi^0) = \sup_{r \in \mathbb{R}^d} (\mu(f_r) - \Lambda_{\pi^0}(f_r)). \quad (41)$$

The first order condition for optimality is given by,

$$g(\mu, r) = 0 \quad (42)$$

where g is the vector valued function that defines the gradient of the objective function in (41):

$$\begin{aligned} g(\mu, r) &:= \nabla_r(\mu(f_r) - \Lambda_{\pi^0}(f_r)) \\ &= \mu(\nabla_r f_r) - \frac{\pi^0(e^{f_r} \nabla_r f_r)}{\pi^0(e^{f_r})} \end{aligned} \quad (43)$$

The derivative of $g(\mu, r)$ with respect to r is given by

$$\begin{aligned} \nabla_r g(\mu, r) &= \mu(\nabla_r^2 f_r) - \left[\frac{\pi^0(e^{f_r} \nabla_r f_r \nabla_r f_r^T) + \pi^0(e^{f_r} \nabla_r^2 f_r)}{\pi^0(e^{f_r})} \right. \\ &\quad \left. - \frac{\pi^0(e^{f_r} \nabla_r f_r) \pi^0(e^{f_r} \nabla_r f_r^T)}{(\pi^0(e^{f_r}))^2} \right] \end{aligned} \quad (44)$$

In these formulae we have extended the definition of $\mu(M)$ for matrix-valued functions M on Z via $[\mu(M)]_{ij} := \mu(M_{ij}) = \sum_z M_{ij}(z) \mu(z)$. On letting $\psi^r = \nabla_r f_r$ we obtain,

$$g(\mu, r) = \mu(\psi^r) - \tilde{\pi}^r(\psi^r) \quad (45)$$

$$\begin{aligned} \nabla_r g(\mu, r) &= \mu(\nabla_r^2 f_r) - \tilde{\pi}^r(\nabla_r^2 f_r) - \\ &\quad [\tilde{\pi}^r(\psi^r \psi^{rT}) - \tilde{\pi}^r(\psi^r) \tilde{\pi}^r(\psi^{rT})] \end{aligned} \quad (46)$$

where the definition of the twisted distribution is as given in (10):

$$\tilde{\pi}^r := \pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)).$$

Proof of Theorem III.2: Without loss of generality, we assume that π^0 has full support over Z . Suppose that the observation sequence \mathbf{Z} is drawn i.i.d. with marginal distribution $\pi \in \mathcal{P}(Z)$. We have $D^{\text{MM}}(\Gamma^n \|\pi^0) \xrightarrow[n \rightarrow \infty]{a.s.} D^{\text{MM}}(\pi \|\pi^0)$ by the law of large numbers.

1) *Proof of part (i):* We first prove the results concerning the bias and variance of the mismatched test statistic. We apply Lemma III.5 to the function $h(\mu) := D^{\text{MM}}(\mu \|\pi^0)$. The other terms appearing in the lemma are taken to be $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \dots, \mathbb{I}_{z_N}(Z_i))^T$, $\mathbf{X} = \mathcal{P}(Z)$, $\bar{x} = \pi^0$, and $S^n = \Gamma^n$. Let $\Xi = \text{Cov}(X^1)$. It is easy to see that $\Xi = \text{diag}(\pi^0) - \pi^0 \pi^{0T}$ and $\Sigma_{\pi^0} = \Psi \Xi \Psi^T$, where Σ_{π^0} is defined in (30), and Ψ is a $d \times N$ matrix defined by,

$$\Psi(i, j) = \psi_i(z_j). \quad (47)$$

This can be expressed as the concatenation of column vectors via $\Psi = [\psi(z_1), \psi(z_2), \dots, \psi(z_N)]$.

We first demonstrate that

$$M = \nabla^2 h(\pi_0) = \Psi^T (\Sigma_{\pi^0})^{-1} \Psi, \quad (48)$$

and then check to make sure that the other requirements of Lemma III.5 are satisfied. The first two conclusions of Theorem III.2 (i) will then follow from Lemma III.5, since

$$\text{trace}(M\Xi) = \text{trace}((\Sigma_{\pi^0})^{-1}\Psi\Xi\Psi^T) = \text{trace}(I_d) = d,$$

and similarly $\text{trace}(M\Xi M\Xi) = \text{trace}(I_d) = d$.

We first prove that under the assumptions of Theorem III.2 (i), there is a function $r : \mathcal{P}(Z) \mapsto \mathbb{R}$ that is C^1 in a neighborhood of π^0 such that $r(\mu)$ solves (41) for μ in this neighborhood. Under the uniqueness assumption (A2), the function $r(\mu)$ coincides with the function given in (A2).

By the assumptions, we know that when $\mu = \pi^0$, (42) is satisfied by r^* with $f_{r^*} \equiv 0$. It follows that $\pi^0 = \tilde{\pi}^{r^*}$. Substituting this into (46), we obtain $\nabla_r g(\mu, r) \Big|_{\substack{\mu=\pi^0 \\ r=r^*}} = -\Sigma_{\pi^0}$, which is negative-definite by Assumption (A3) and Lemma III.1. Therefore, by the Implicit Function Theorem, there is an open neighborhood U around $\mu = \pi^0$, an open neighborhood V of r^* , and a continuously differentiable function $r : U \rightarrow V$ that satisfies $g(\mu, r(\mu)) = 0$, for $\mu \in U$. This fact together with Assumptions (A2) and (A3) ensure that when $\mu \in U \cap B$, the vector $r(\mu)$ uniquely achieves the supremum in (41).

Taking the total derivative of (42) with respect to $\mu(z)$ we get,

$$\frac{\partial r(\mu)}{\partial \mu(z)} = - \left[\nabla_r g(\mu, r(\mu)) \right]^{-1} \frac{\partial g(\mu, r(\mu))}{\partial \mu(z)}. \quad (49)$$

Consequently, when $\mu = \pi^0$,

$$\frac{\partial r(\mu)}{\partial \mu(z)} \Big|_{\mu=\pi^0} = \Sigma_{\pi^0}^{-1} \psi(z). \quad (50)$$

These results enable us to identify the first and second order derivative of $h(\mu) = D^{\text{MM}}(\mu \| \pi^0)$. Applying $g(\mu, r(\mu)) = 0$, we obtain the derivatives of h as follows,

$$\frac{\partial}{\partial \mu(z)} h(\mu) = f_{r(\mu)}(z). \quad (51)$$

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\mu) = (\nabla_r f_{r(\mu)}(z))^T \frac{\partial r(\mu)}{\partial \mu(\bar{z})}. \quad (52)$$

When $\mu = \pi^0$, substituting (50) in (52), we obtain (48).

We now verify the remaining conditions required for applying Lemma III.5:

- (a) It is straightforward to see that $h(\pi^0) = 0$.
- (b) The function h is uniformly bounded since $h(\mu) = D^{\text{MM}}(\mu \| \pi^0) \leq D(\mu \| \pi^0) \leq \max_z \log(\frac{1}{\pi^0(z)})$ and π^0 has full support.
- (c) Since $f_{r(\mu)} = 0$ when $\mu = \pi^0$, it follows by (51) that $\frac{\partial}{\partial \mu(z)} h(\mu) \Big|_{\mu=\pi^0} = 0$.

(d) Pick a compact $K \subset U \cap B$ so that K contains π^0 as a relative interior point, and $K \subset \{\mu \in \mathcal{P}(Z) : \max_u |\mu(u) - \pi^0(u)| < \frac{1}{2} \min_u |\pi^0(u)|\}$. This choice of K ensures that $\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\{S^n \notin K\} > 0$. Note that since $r(\mu)$ is continuously differentiable on $U \cap B$, it follows by (51) and (52) that h is C^2 on K .

Thus the results on convergence of the bias and variance follow from Lemma III.5.

The weak convergence result is proved using Lemma III.6 and Lemma III.7. We observe that the covariance matrix of the Gaussian vector W given in Lemma III.6 is $\Sigma_W = \Xi = \text{diag}(\pi^0) - \pi^0 \pi^{0T}$. This does not have full rank since $\Xi \mathbf{1} = 0$, where $\mathbf{1}$ is the $N \times 1$ vector of ones. Hence we can write,

$$\Xi = GG^T$$

where G is an $N \times k$ matrix for some $k < N$. In fact, since the support of π^0 is full, we have $k = N - 1$ (see Lemma III.1). Based on this representation we can write $W = GV$, where $V \sim \mathcal{N}(0, I_k)$.

Now, by Lemma III.6, the limiting random variable is given by $U := \frac{1}{2} W^T M W = \frac{1}{2} V^T G^T M G V$, where $M = \nabla_{\mu}^2 D^{\text{MM}}(\mu \| \pi^0) \Big|_{\pi^0} = \Psi^T (\Psi \Xi \Psi^T)^{-1} \Psi$. We observe that the matrix $D = G^T M G$ satisfies $D^2 = D$. Moreover, since $\Psi \Xi \Psi^T$ has rank d under Assumption (A3), matrix D also has rank d . Applying Lemma III.7 to matrix D , we conclude that $U \sim \frac{1}{2} \chi_d^2$.

2) *Proof of part (ii):* The conclusion of part (ii) is derived using part (i) and the decomposition in (38). We will study the bias, variance, and limiting distribution of each term in the decomposition.

For the second term, note that the dimensionality of the function class \mathcal{G} is also d . Applying part (i) of this theorem to $D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)$, we conclude that its asymptotic bias and variance are given by

$$\lim_{n \rightarrow \infty} \mathbb{E}[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)] = \frac{1}{2} d, \quad (53)$$

$$\lim_{n \rightarrow \infty} \text{Var}[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)] = \frac{1}{2} d. \quad (54)$$

For the third term, since \mathbf{Z} is i.i.d. with marginal π^1 , we have

$$\mathbb{E}[\langle \Gamma^n - \pi^1, f_{r^*} \rangle] = 0, \quad (55)$$

$$\text{Var}[n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle] = \text{Cov}_{\pi^1}(f_{r^*}). \quad (56)$$

The bias result (34) follows by combining (53), (55) and using the decomposition (38). To prove the variance result (35), we again apply the decomposition (38) to obtain,

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Var}[n^{\frac{1}{2}} D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \| \pi^0)] &= \lim_{n \rightarrow \infty} \{ \text{Var}[n^{\frac{1}{2}} D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)] + \text{Var}[n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle] \\ &\quad + 2E[n^{\frac{1}{2}} (D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1) - \mathbb{E}[D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)]) \\ &\quad \quad n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle] \}. \end{aligned} \quad (57)$$

From (54) it follows that the limiting value of the first term on the right hand side of (57) is 0. The limiting value of the third term is also 0 by applying the Cauchy-Bunyakovsky–Schwarz inequality. Thus, (57) together with (56) gives (35).

Finally, we prove the weak convergence result (36) by again applying the decomposition (38). By (53) and (54), we conclude that the second term $n^{\frac{1}{2}} D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \|\pi^1)$ converges in mean square to 0 as $n \rightarrow \infty$. The weak convergence of the third term is given in (39). Applying Slutsky's theorem, we obtain (36). ■

Proof of Theorem III.3: The proof of this result is very similar to that of Theorem III.2 (ii) except that we use the decomposition in (37) with $\mu = \Gamma^n$. We first prove the following generalizations of (53) and (54) that characterizes the asymptotic mean and variance of the second term in (37) with $\mu = \Gamma^n$:

$$\lim_{n \rightarrow \infty} \mathbb{E}[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \|\tilde{\pi})] = \frac{1}{2} \text{trace}(\Sigma_{\pi^1} (\Sigma_{\tilde{\pi}})^{-1}) \quad (58)$$

$$\lim_{n \rightarrow \infty} \text{Var}[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \|\tilde{\pi})] = \frac{1}{2} \text{trace}(\Sigma_{\pi^1} (\Sigma_{\tilde{\pi}})^{-1} \Sigma_{\pi^1} (\Sigma_{\tilde{\pi}})^{-1}) \quad (59)$$

where $\mathcal{G} = \mathcal{F} - f_{r^1}$, and $\tilde{\pi}$ is defined in the statement of the proposition. The argument is similar to that of Theorem III.2 (i): We denote $\tilde{f}_r := f_r - f_{r^1}$, and define $h(\mu) := D_{\mathcal{G}}^{\text{MM}}(\mu \|\tilde{\pi}) = \sup_{r \in \mathbb{R}^d} (\mu(\tilde{f}_r) - \Lambda_{\tilde{\pi}}(\tilde{f}_r))$.

To apply Lemma III.5, we prove the following

$$h(\pi^1) = 0, \quad (60)$$

$$\nabla_{\mu} h(\pi^1) = 0, \quad (61)$$

$$\text{and} \quad M = \nabla_{\mu}^2 h(\pi^1) = \Psi^T (\Sigma_{\tilde{\pi}})^{-1} \Psi. \quad (62)$$

The last two inequalities (61) and (62) are analogous to (51) and (52). We can also verify that the rest of the conditions of Lemma III.5 hold. This establishes (58) and (59).

To prove (60), first note that the supremum in the optimization problem defining $D^{\text{MM}}(\pi^1 \|\tilde{\pi})$ is achieved by \tilde{f}_{r^1} , and we know by definition that $\tilde{f}_{r^1} = 0$. Together with the definition $D^{\text{MM}}(\pi^1 \|\tilde{\pi}) = \pi^1(\tilde{f}_{r^1}) - \Lambda_{\tilde{\pi}}(\tilde{f}_{r^1})$, we obtain (60).

Redefine $g(\mu, r) := \nabla_r (\mu(\tilde{f}_r) - \Lambda_{\tilde{\pi}}(\tilde{f}_r))$. The first order optimality condition of the optimization problem defining $D^{\text{MM}}(\mu \|\tilde{\pi})$ gives $g(\mu, r) = 0$. The assumption that \mathcal{F} is a linear function class implies that \tilde{f}_r is linear in r . Consequently $\nabla_r^2 \tilde{f}_r = 0$. By the same argument that leads to (44), we can show that

$$\nabla_r g(\mu, r) = - \left[\frac{\tilde{\pi}(e^{\tilde{f}_r} \nabla_r \tilde{f}_r \nabla_r \tilde{f}_r^T)}{\tilde{\pi}(e^{\tilde{f}_r})} - \frac{\tilde{\pi}(e^{\tilde{f}_r} \nabla_r \tilde{f}_r) \tilde{\pi}(e^{\tilde{f}_r} \nabla_r \tilde{f}_r^T)}{(\tilde{\pi}(e^{\tilde{f}_r}))^2} \right] \quad (63)$$

Together with the fact that $\tilde{f}_{r^1} = 0$ and $\nabla_r \tilde{f}_r = \nabla_r f_r$, we obtain

$$\nabla_r g(\mu, r) \Big|_{\substack{\mu = \pi^1 \\ r = r^1}} = -\Sigma_{\tilde{\pi}}. \quad (64)$$

Proceeding as in the proof of Theorem III.2 (i), we obtain (61) and (62).

Now using similar steps as in the proof of Theorem III.2 (ii), and noticing that $\log(\frac{\tilde{\pi}}{\pi_0}) = f_{r^1}$, we can establish the following results on the third term of (37):

$$\begin{aligned} \mathbb{E}[\langle \Gamma^n - \pi^1, \log(\frac{\tilde{\pi}}{\pi_0}) \rangle] &= 0 \\ \text{Var}[n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, \log(\frac{\tilde{\pi}}{\pi_0}) \rangle] &= \text{Cov}_{\pi^1}(f_{r^1}) \\ n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, \log(\frac{\tilde{\pi}}{\pi_0}) \rangle &\xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \text{Cov}_{\pi^1}(f_{r^1})). \end{aligned}$$

Continuing the same arguments as in Theorem III.2 (i), we obtain the result of Theorem III.3. ■

A. Interpretation of the asymptotic results and performance comparison

The asymptotic results established above can be used to study the finite sample performance of the mismatched test and Hoeffding test. Recall that in the discussion surrounding Figure 1 we concluded that the approximation obtained from a Central Limit Theorem gives much better estimates of error probabilities as compared to those suggested by Sanov's theorem.

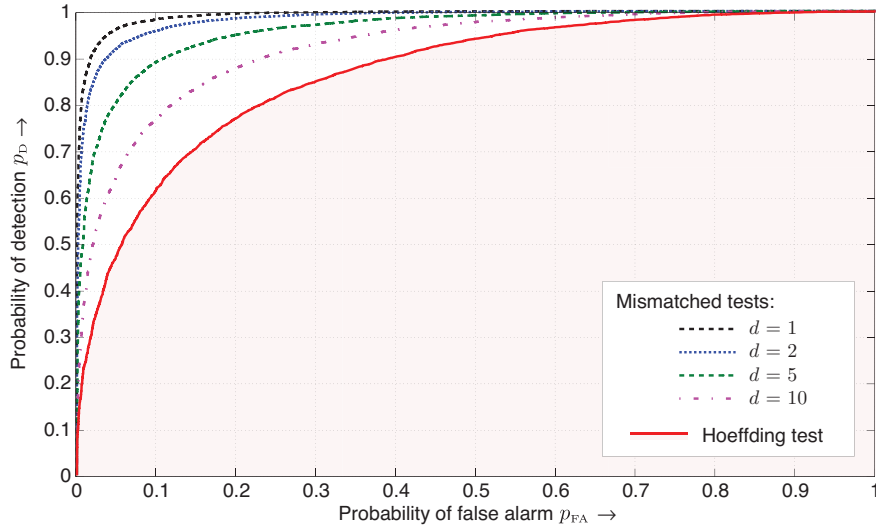


Fig. 4. Comparison of ROCs of Hoeffding and mismatched tests.

Suppose the log-likelihood ratio function $\log(\pi^1/\pi^0)$ lies in the function class \mathcal{F} . In this case, the results of Theorem III.2 and Lemma III.4 are informally summarized in the following approximations:

With Γ^n denoting the empirical distributions of the i.i.d. process Z ,

$$D^{\text{MM}}(\Gamma^n \parallel \pi^0) \approx \begin{cases} D(\pi^0 \parallel \pi^0) + \frac{1}{2} \frac{1}{n} \sum_{k=1}^d W_k^2, & Z_i \sim \pi^0 \\ D(\pi^1 \parallel \pi^0) + \frac{1}{2} \frac{1}{n} \sum_{k=1}^d W_k^2 + \frac{1}{\sqrt{n}} \sigma_1 U, & Z_i \sim \pi^1 \end{cases} \quad (65)$$

where $\{W_k\}$ is i.i.d., $N(0, 1)$, and U is also $N(0, 1)$ but not independent of the W_k 's. The standard deviation σ_1 is given in Theorem III.2. These distributional approximations are valid for large n , and are subject to assumptions on the function class used in the theorem.

We observe from (65) that, for large enough n , when the observations are drawn under π^0 , the mismatched divergence is well approximated by $\frac{1}{2n}$ times a chi-squared random variable with d degrees of freedom. We also observe that when the observations are drawn under π^1 , the mismatched divergence is well approximated by a Gaussian random variable with mean $D(\pi^1 \parallel \pi^0)$ and with a variance proportional to $\frac{1}{n}$ and independent of d . Since the mismatched test can be interpreted as a GLRT, these results capture the rate of degradation of the finite sample performance of a GLRT as the dimensionality of the parameterized family of alternate hypotheses increases. We corroborate this intuitive reasoning through Monte Carlo simulation experiments.

We estimated via simulation the performance of the Hoeffding test and mismatched tests designed using a linear function class. We compared the error probabilities of these tests for an alphabet size of $N = 19$ and sequence length of $n = 40$. We chose π^0 to be the uniform distribution, and π^1 to be the distribution obtained by convolving two uniform distributions on sets of size $(N + 1)/2$. We chose the basis function ψ_1 appearing in (26) to be the log-likelihood ratio between π^1 and π^0 , viz.,

$$\psi_1(z_i) = \log \frac{\pi^1(z_i)}{\pi^0(z_i)}, \quad 1 \leq i \leq N$$

and the other basis functions $\psi_2, \psi_3, \dots, \psi_d$ were chosen uniformly at random. Figure 4 shows a comparison of the ROCs of the Hoeffding test and mismatched tests for different values of dimension d . Plotted on the x -axis is the probability of false alarm, i.e., the probability of misclassification under π^0 ; shown on the y -axis is the probability of detection, i.e., the probability of correct classification under π^1 . The various points on each ROC curve are obtained by varying the threshold η used in the Hoeffding test of (4) and mismatched test of (19).

From Figure 4 we see that as d increases the performance of the mismatched tests degrades. This is consistent with the approximation (65) which suggests that the variance of the mismatched divergence increases with d . Furthermore, as we saw earlier, the Hoeffding test can be interpreted as a special case of

the mismatched test for a specific choice of the function class with $d = N - 1$ and hence the performance of the mismatched test matches the performance of the Hoeffding test when $d = N - 1$.

To summarize, the above results suggest that although the Hoeffding test is optimal in an error-exponent sense, it is disadvantageous in terms of finite sample error probabilities to blindly use the Hoeffding test if it is known a priori that the alternate distribution belongs to some parameterized family of distributions.

IV. CONCLUSIONS

The mismatched test provides a solution to the universal hypothesis testing problem that can incorporate prior knowledge in order to reduce variance. The main results of Section III show that the variance reduction over Hoeffding's optimal test is substantial when the state space is large.

The dimensionality of the function class can be chosen by the designer to ensure that the bias and variance are within tolerable limits. It is in this phase of design that prior knowledge is required to ensure that the error-exponent remains sufficiently large under the alternate hypothesis (see e.g. Corollary II.1). In this way the designer can make effective tradeoffs between the power of the test and the variance of the test statistic.

The mismatched divergence provides a unification of several approaches to robust and universal hypothesis testing. Although constructed in an i.i.d. setting, the mismatched tests are applicable in very general settings, and the performance analysis presented here is easily generalized to any stationary process satisfying the Central Limit Theorem.

There are many directions for future research. Topics of current research include,

- (i) Algorithms for basis synthesis and basis adaptation.
- (ii) Extensions to Markovian models.
- (iii) Extensions to change detection.

Initial progress in basis synthesis is reported in [22]. Recent results addressing the computational complexity of the mismatched test are reported in [20]. Although the exact computation of the mismatched divergence requires the solution of an optimization problem, we describe a computationally tractable approximation in [20]. We are also actively pursuing applications to problems surrounding building energy and surveillance. Some initial progress is reported in [23].

ACKNOWLEDGEMENTS

We thank Prof. Imre Csiszár for his insightful comments.

APPENDIX

A. Excess codelength for source coding with training

The results in Theorem III.2 give us the asymptotic behavior of $D(\Gamma^n \parallel \pi)$ but what we need here is the behavior of $D(\pi \parallel \Gamma^n)$. Define

$$h(\mu) = \begin{cases} D(\pi \parallel \mu) & \text{if } \mu \in \mathbb{P}_{\epsilon/2} \\ D(\pi \parallel \pi^u) & \text{else} \end{cases}.$$

It is clear that h is uniformly bounded from above by $\log \frac{2}{\epsilon}$. Although h is not continuous at the boundary of $\mathbb{P}_{\epsilon/2}$, a modified version of Lemmas III.5 and III.6 can be applied to the function h to establish the results of (16) following the same steps used in proving Theorem III.2. The Hessian matrix M appearing in the statement of the lemmas is given by,

$$M = \nabla^2 h(\pi) = \text{diag}(\pi)^{-1}.$$

Hence, $\text{trace}(M\Omega) = \text{trace}(M\Omega M\Omega) = N - 1$.

B. Proof of Lemma III.4

Proof: In the following chain of identities, the first, third and fifth equalities follow from relation (25) and Proposition II.5.

$$\begin{aligned} D_{\mathcal{F}}^{\text{MM}}(\mu \parallel \pi^0) &= D(\mu \parallel \pi^0) - \inf\{D(\mu \parallel \nu) : \nu = \pi^0 \exp(f - \Lambda_{\pi^0}(f)), f \in \mathcal{F}\} \\ &= D(\mu \parallel \tilde{\pi}) + \langle \mu, \log(\frac{\tilde{\pi}}{\pi^0}) \rangle - \inf\{D(\mu \parallel \nu) : \nu = \tilde{\pi} \exp(f - \Lambda_{\tilde{\pi}}(f)), f \in \mathcal{G}\} \\ &= D_{\mathcal{G}}^{\text{MM}}(\mu \parallel \tilde{\pi}) + \langle \mu, \log(\frac{\tilde{\pi}}{\pi^0}) \rangle \\ &= D_{\mathcal{G}}^{\text{MM}}(\mu \parallel \tilde{\pi}) + \langle \mu - \pi^1, \log(\frac{\tilde{\pi}}{\pi^0}) \rangle + D(\pi^1 \parallel \pi^0) - D(\pi^1 \parallel \tilde{\pi}) \\ &= D_{\mathcal{G}}^{\text{MM}}(\mu \parallel \tilde{\pi}) + \langle \mu - \pi^1, \log(\frac{\tilde{\pi}}{\pi^0}) \rangle + D_{\mathcal{F}}^{\text{MM}}(\pi^1 \parallel \pi^0) \end{aligned}$$

■

C. Proof of Lemma III.5

The following simple lemma will be used in multiple places in the proof that follows.

Lemma A.1. *If a sequence of random variables $\{A^n\}$ satisfies $E[A^n] \xrightarrow[n \rightarrow \infty]{} a$ and $\{E[(A^n)^2]\}$ is a bounded sequence, and another sequence of random variables $\{B^n\}$ satisfies $B^n \xrightarrow[n \rightarrow \infty]{m.s.} b$, then $E[A^n B^n] \xrightarrow[n \rightarrow \infty]{} ab$. □*

Proof of Lemma III.5: Without loss of generality, we can assume that the mean \bar{x} is the origin in \mathbb{R}^m and that $h(\bar{x}) = 0$.

Since the Hessian is continuous over the set K , we have by Taylor's theorem:

$$n(h(S^n) - \nabla h(\bar{x})^T S^n) \mathbb{I}_{\{S^n \in K\}} = n[h(\bar{x}) + \frac{1}{2} S^{nT} \nabla^2 h(\tilde{S}^n) S^n] \mathbb{I}_{\{S^n \in K\}} \quad (66)$$

$$= \frac{n}{2} S^{nT} \nabla^2 h(\tilde{S}^n) S^n \mathbb{I}_{\{S^n \in K\}} \quad (67)$$

where $\tilde{S}^n = \gamma S^n$ for some $\gamma = \gamma(n) \in [0, 1]$. By the strong law of large numbers we have $S^n \xrightarrow[n \rightarrow \infty]{a.s.} \bar{x}$. Hence $\tilde{S}^n \xrightarrow[n \rightarrow \infty]{a.s.} \bar{x}$ and $\nabla^2 h(\tilde{S}^n) \xrightarrow[n \rightarrow \infty]{a.s.} \nabla^2 h(\bar{x}) = M$ since $\nabla^2 h$ is continuous at \bar{x} . Now by the boundedness of the second derivative over K and the fact that

$$\mathbb{I}_{\{S^n \in K\}} \xrightarrow[n \rightarrow \infty]{a.s.} 1$$

we have $(\nabla^2 h(\tilde{S}^n))_{i,j} \mathbb{I}_{\{S^n \in K\}} \xrightarrow[n \rightarrow \infty]{m.s.} M_{i,j}$.

Under the assumption that \mathbf{X} is i.i.d. on the compact set \mathbf{X} , we have

$$\mathbb{E}[n S_i^n S_j^n] = \Sigma_{i,j} \text{ for all } n,$$

and $\mathbb{E}[(n S_i^n S_j^n)^2]$ converges to a finite quantity as $n \rightarrow \infty$. Hence the results of Lemma A.1 are applicable with $A^n = n S_i^n S_j^n$ and $B^n = \nabla^2 h(\tilde{S}^n)_{i,j} \mathbb{I}_{\{S^n \in K\}}$, which gives:

$$\mathbb{E}[n S_i^n S_j^n \nabla^2 h(\tilde{S}^n)_{i,j} \mathbb{I}_{\{S^n \in K\}}] \xrightarrow[n \rightarrow \infty]{} \Sigma_{i,j} M_{i,j}. \quad (68)$$

Thus we have

$$\begin{aligned} \mathbb{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n) \mathbb{I}_{\{S^n \in K\}}] &= \mathbb{E}[\frac{n}{2} S^{nT} \nabla^2 h(\tilde{S}^n) S^n \mathbb{I}_{\{S^n \in K\}}] \\ &\xrightarrow[n \rightarrow \infty]{} \frac{1}{2} \text{trace}(M \Xi). \end{aligned} \quad (69)$$

Since \mathbf{X} is compact, h is continuous, and h is differentiable at \bar{x} , it follows that there are scalars \bar{h} and \bar{x} such that $\sup_{x \in \mathbf{X}} |h(x)| \leq \bar{h}$ and $|\nabla h(\bar{x})^T S^n| < \bar{x}$. Hence,

$$|\mathbb{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n) \mathbb{I}_{\{S^n \notin K\}}]| \leq n(\bar{h} + \bar{x}) \mathbb{P}\{S^n \notin K\} \xrightarrow[n \rightarrow \infty]{} 0 \quad (70)$$

where we use the assumption that the $\mathbb{P}\{S^n \notin K\}$ decays exponentially in n . Combining (69) and (70) and using the fact that S^n has zero mean, we have

$$\mathbb{E}[nh(S^n)] = \mathbb{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)] \xrightarrow[n \rightarrow \infty]{} \frac{1}{2} \text{trace}(M \Xi).$$

This establishes the result of (i).

Under the condition that the directional derivative is zero, (67) can be written as

$$nh(S^n)\mathbb{I}_{\{S^n \in K\}} = \frac{n}{2} S^{nT} \nabla^2 h(\tilde{S}^n) S^n \mathbb{I}_{\{S^n \in K\}}. \quad (71)$$

Now by squaring (71), we have

$$(nh(S^n)\mathbb{I}_{\{S^n \in K\}})^2 = \frac{n^2}{4} \sum_{i,j,k,\ell} \left[S_i^n (\nabla^2 h(\tilde{S}^n))_{i,j} S_j^n S_k^n (\nabla^2 h(\tilde{S}^n))_{k,\ell} S_\ell^n \mathbb{I}_{\{S^n \in K\}} \right].$$

As before, by the boundedness of the Hessian we have:

$$(\nabla^2 h(\tilde{S}^n))_{i,j} (\nabla^2 h(\tilde{S}^n))_{k,\ell} \mathbb{I}_{\{S^n \in K\}} \xrightarrow[n \rightarrow \infty]{m.s.} M_{i,j} M_{k,\ell}$$

It can also be shown that

$$\mathbb{E}[n^2 S_i^n S_j^n S_k^n S_\ell^n] = \frac{F_{i,j,k,\ell}}{n} + \Sigma_{i,j} \Sigma_{k,\ell} + \Sigma_{j,k} \Sigma_{i,\ell} + \Sigma_{i,k} \Sigma_{j,\ell} \text{ for all } n$$

where $F_{i,j,k,\ell} = \mathbb{E}[X_i^1 X_j^1 X_k^1 X_\ell^1]$. Moreover, $\mathbb{E}[(n^2 S_i^n S_j^n S_k^n S_\ell^n)^2]$ is finite for each n and converges to a finite quantity as $n \rightarrow \infty$ since the moments of X^i are finite. Thus we can again apply Lemma A.1 to see that

$$\begin{aligned} & \mathbb{E}[n^2 S_i^n \nabla^2 h(\tilde{S}^n)_{i,j} S_j^n S_k^n \nabla^2 h(\tilde{S}^n)_{k,\ell} S_\ell^n \mathbb{I}_{\{S^n \in K\}}] \\ & \xrightarrow[n \rightarrow \infty]{} (\Sigma_{i,j} \Sigma_{k,\ell} + \Sigma_{j,k} \Sigma_{i,\ell} + \Sigma_{i,k} \Sigma_{j,\ell}) M_{i,j} M_{k,\ell}. \end{aligned}$$

Putting together terms and using (71) we obtain:

$$\mathbb{E}[(nh(S^n))^2 \mathbb{I}_{\{S^n \in K\}}] \xrightarrow[n \rightarrow \infty]{} \frac{1}{2} \text{trace}(M \Xi M \Xi) + \frac{1}{4} (\text{trace}(M \Xi))^2.$$

Now similar to (70) we have:

$$|\mathbb{E}[(nh(S^n))^2 \mathbb{I}_{\{S^n \notin K\}}]| \leq n^2 \bar{h}^2 \mathbb{P}\{S^n \notin K\} \xrightarrow[n \rightarrow \infty]{} 0. \quad (72)$$

Consequently

$$\mathbb{E}[(nh(S^n))^2] \xrightarrow[n \rightarrow \infty]{} \frac{1}{2} \text{trace}(M \Xi M \Xi) + \frac{1}{4} (\text{trace}(M \Xi))^2$$

which gives (ii). ■

D. Proof of Lemma III.6

We know from (2) that Γ^n can be written as an empirical average of i.i.d. vectors. Hence, it satisfies the central limit theorem which says that,

$$n^{\frac{1}{2}}(\Gamma^n - \mu) \xrightarrow[n \rightarrow \infty]{d.} W \quad (73)$$

where the distribution of W is defined below (40).

Considering a second-order Taylor's expansion and using the condition on the directional derivative, we have,

$$n(h(\Gamma^n) - h(\mu)) = \frac{1}{2}n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu))$$

where $\tilde{\Gamma}^n = \gamma\Gamma^n + (1 - \gamma)\mu$ for some $\gamma = \gamma(n) \in [0, 1]$. We also know by the strong law of large numbers that Γ^n and hence $\tilde{\Gamma}^n$ converge to μ almost surely. By the continuity of the Hessian, we have

$$\nabla^2 h(\tilde{\Gamma}^n) \xrightarrow[n \rightarrow \infty]{a.s.} \nabla^2 h(\mu). \quad (74)$$

By applying the vector-version of Slutsky's theorem [24], together with (73) and (74), we conclude

$$n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu)) \xrightarrow[n \rightarrow \infty]{d.} \frac{1}{2}W^T \nabla^2 h(\mu)W,$$

thus establishing the lemma.

E. Proof of Lemma III.7

Proof: The assumption that D is a projection matrix implies that $D^2 = D$. Let $\{u^1, \dots, u^m\}$ denote an orthonormal basis, chosen so that the first K vectors span the range space of D . Hence $Du^i = u^i$ for $1 \leq i \leq K$, and $Du^i = 0$ for all other i .

Let U denote the unitary matrix whose m columns are $\{u^1, \dots, u^m\}$. Then $\tilde{V} = UV$ is also an $\mathcal{N}(0, I_m)$ random variable, and hence DV and $D\tilde{V}$ have the same Gaussian distribution.

To complete the proof we demonstrate that $\|D\tilde{V}\|^2$ has a chi-squared distribution: By construction the vector $\tilde{Y} = D\tilde{V}$ has components given by

$$\tilde{Y}_i = \begin{cases} \tilde{V}_i & 1 \leq i \leq K \\ 0 & K < i \leq m \end{cases}$$

It follows that $\|\tilde{Y}\|^2 = \|D\tilde{V}\|^2 = \tilde{V}_1^2 + \dots + \tilde{V}_K^2$ has a chi-squared distribution with K degrees of freedom. ■

REFERENCES

- [1] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, "Statistical SVMs for robust detection, supervised learning, and universal classification," in *Proc. of IEEE Information Theory Workshop on Networking and Information Theory*, Volos, Greece, 2009, pp. 62–66.
- [2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.
- [3] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.
- [4] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2215–2229, 2002.
- [5] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep 1996.
- [6] E. Abbe, M. Medard, S. Meyn, and L. Zheng, "Finding the best mismatched detector for channel coding and hypothesis testing," *Information Theory and Applications Workshop, 2007*, pp. 284–288, 29 2007-Feb. 2 2007.
- [7] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.
- [8] C. Pandit, S. Meyn, and V. Veeravalli, "Asymptotic robust Neyman-Pearson hypothesis testing based on moment classes," in *Proc. of IEEE International Symposium on Information Theory*, Chicago, 2004, p. 220.
- [9] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stoch. Proc. Applns.*, vol. 116, no. 5, pp. 724–756, May 2006.
- [10] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. I. II," *Comm. Pure Appl. Math.*, vol. 28, pp. 1–47; *ibid.* **28** (1975), 279–301, 1975.
- [11] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, 2010, to appear.
- [12] B. Clarke and A. R. Barron, "Information theoretic asymptotics of bayes' methods," Univ. of Illinois, Department of Statistics, Tech. Rep. 26, July 1989.
- [13] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, 1990.
- [14] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statistics*, vol. 9, pp. 60–62, 1938.
- [15] M. Iltis, "Sharp asymptotics of large deviations in \mathbb{R}^d ," *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, 1995. [Online]. Available: <http://www.springerlink.com/content/3768273345028TR4>
- [16] P. Harremoës, "Testing goodness-of-fit via rate distortion," in *Proc. of IEEE Information Theory Workshop on Networking and Information Theory*, Volos, Greece, June 2009, pp. 17–21.
- [17] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 285–290, 1991.
- [18] —, "Correction to: "On universal hypotheses testing via large deviations"," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, part 1, p. 698, 1991.
- [19] Z. Enlu, M. C. Fu, and S. I. Marcus, "Solving continuous-state pomdps via density projection," *IEEE Trans. Autom. Control*, vol. 55, no. 5, pp. 1101 – 1116, May 2010.

- [20] J. Unnikrishnan, “Decision-making under statistical uncertainty,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL, August 2010.
- [21] J. Unnikrishnan, S. Meyn, and V. V. Veeravalli, “On thresholds for goodness of fit tests,” presented at IEEE Information Theory Workshop, Dublin, 2010. [Online]. Available: [http://www.ifp.illinois.edu/~sim\\$junnikr2/pdfs/ITW10.pdf](http://www.ifp.illinois.edu/~sim$junnikr2/pdfs/ITW10.pdf)
- [22] D. Huang and S. Meyn, “Feature extraction for universal hypothesis testing via rank-constrained optimization,” in *Information Theory Proceedings (ISIT)*, 2010 IEEE International Symposium on, Jun. 2010, pp. 1618 – 1622.
- [23] S. Meyn, A. Surana, Y. Lin, and S. Narayanan, “Anomaly detection using projective Markov models in a distributed sensor network,” in *Proc. of 48th IEEE Conference on Decision and Control*, Shanghai, December 2009, pp. 4662–4669.
- [24] P. Billingsley, *Convergence of Probability Measures*. New York: John Wiley & Sons, 1968.